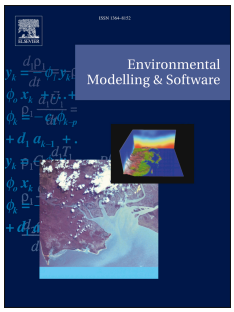
A novel ensemble-based conceptual-data-driven approach for improved streamflow simulations

Sikorska-Senoner, Anna E ; Quilty, John M

Abstract: A novel framework, an ensemble-based conceptual-data-driven approach (CDDA), is developed that integrates a hydrological model (HM) with a data-driven model (DDM) to simulate an ensemble of HM residuals. Thus, a CDDA delivers an ensemble of 'residual-corrected' streamflow simulations. This framework is beneficial because it respects hydrological processes via the HM and it profits from the DDM's ability to simulate the complex relationship between residuals and input variables. The CDDA enables exploring different DDMs to identify the most suitable model. Eight DDMs are explored: Multiple Linear Regression (MLR), k Nearest Neighbours Regression (kNN), Second-Order Volterra Series Model, Artificial Neural Networks (ANN), and two variants of eXtreme Gradient Boosting (XGB) and Random Forests (RF). The proposed CDDA, tested on three Swiss catchments, was able to improve the mean continuous ranked probability score by 16-29% over the standalone HM. Based on these results, XGB and RF are recommended for simulating the HM residuals.

DOI: https://doi.org/10.1016/j.envsoft.2021.105094

Journal Pre-proof



A novel ensemble-based conceptual-data-driven approach for improved streamflow simulations

Anna E. Sikorska-Senoner, John M. Quilty

PII: S1364-8152(21)00137-7

DOI: https://doi.org/10.1016/j.envsoft.2021.105094

Reference: ENSO 105094

To appear in: *Environmental Modelling and Software*

Accepted Date: 24 May 2021

Please cite this article as: Sikorska-Senoner, A.E., Quilty, J.M., A novel ensemble-based conceptual-data-driven approach for improved streamflow simulations, *Environmental Modelling and Software*, https://doi.org/10.1016/j.envsoft.2021.105094.

A novel ensemble-based conceptual-data-driven approach for improved streamflow simulations

Anna E. Sikorska-Senoner[1] and John M. Quilty[2]

[1]Department of Geography, University of Zurich, Winterthurerstrasse 190, 8057 Zürich, Switzerland

[2]Department of Civil and Environmental Engineering, University of Waterloo, 200 University Avenue West, Waterloo, ON, N2L 3G1, Canada

Corresponding author:
Anna E. Sikorska-Senoner: e-mail address: anna.senoner@geo.uzh.ch, Telephone: 0041 (0)44 6356532

Highlights:

- Conceptual-data-driven approach (CDDA) proposed for ensemble streamflow simulation
- The CDDA couples a data-driven model (DDM) and a hydrological model (HM)
- Eight DDMs are explored as a potential predictor of the HM residual ensemble
- CDDA improves the mean continuous ranked probability score vs. standalone HM
- eXtreme Gradient Boosting and Random Forests are recommended to modelHM residuals

Authors ORCIDs:

Anna E. Sikorska-Senoner: 0000-0002-5273-1038

John M. Quilty: 0000-0002-0207-8077

Abstract

A novel framework, an ensemble-based conceptual-data-driven approach (CDDA), is developed that integrates a hydrological model (HM) with a data-driven model (DDM) to simulate an ensemble of HM residuals. Thus, a CDDA delivers an ensemble of 'residual-corrected' streamflow simulations. This framework is beneficial because it respects hydrological processes via the HM and it profits from the DDM's ability to simulate the complex relationship between residuals and input variables. The CDDA enables exploring different DDMs to identify the most suitable model. Eight DDMs are explored: Multiple Linear Regression (MLR), k Nearest Neighbours Regression (kNN), Second-Order Volterra Series Model, Artificial Neural Networks (ANN), and two variants of eXtreme Gradient Boosting (XGB) and Random Forests (RF). The proposed CDDA, tested on three Swiss catchments, was able to improve the mean continuous ranked probability score by 16-29% over the standalone HM. Based on these results, XGB and RF are recommended for simulating the HM residuals.

Keywords:

Ensemble, streamflow simulation, data-driven model, hydrological model, residuals, uncertainty

1 Introduction

The water resources domain has witnessed growing interest in using data-driven models (DDMs) to improve precipitation-runoff simulation (Shen, 2018; Nearing et al., 2020), whether by directly replacing (e.g., Kratzert et al., 2019) or being used in conjunction with (e.g., Boucher et al., 2020) hydrological models (HMs). In this study, the use of DDMs are explored for improving ensemble-based streamflow simulations generated by an ensemble of HMs, where the DDMs act as 'residual' models. In other words, DDMs are used to correct the model residuals (errors) stemming from an ensemble of streamflow simulations generated by an HM.

HM errors impair estimates of design floods, risk assessment, and, in general, water resources management (Farmer & Vogel, 2016). In general, HM errors result from numerous sources of uncertainty such as model structure, input and output variables, parameters, initial conditions, etc. (Kuczera et al., 2010; Renard et al., 2011; McMillan et al., 2012; Sikorska & Renard, 2017). Although HM errors have been traditionally assumed to be independent and normally distributed, it is now widely accepted that this assumption does not hold since HM

residuals are highly auto-correlated (Sorooshian & Dracup, 1980; Yang et al., 2007; Sikorska et al., 2012). Hence, they cannot be modelled as a normally-distributed random process. As a result, and alternative methods that do not rely on these limiting assumptions have gained widespread attention in hydrology.

Available methods to tackle errors in HM can be generally classified into two major groups (Ni et al., 2020). The first group considers a classical HM and uses various tools for model error identification and estimating related uncertainties, from methodologically simple Monte Carlo experiments with different HM parameter sets (e.g., Padiyedath Gopalan et al. (2019)) to more advanced likelihood-based methods (see examples below). The second group uses data-driven approaches to directly link the HM simulations to its input variables within a hybrid modelling approach (Althoff et al. 2021).

Among the first group of methods, Bayesian-based methods have received the most attention since they are able to simulate the auto-correlation of model residuals and quantify contributions from different uncertainty sources (Mantovan & Todini, 2006, Renard et al., 2011). Yet, they require making an explicit assumption on model error properties (with introducing additional parameters for the model inference) and a definition of the likelihood function to sample from the posterior distribution (Montanari & Koutsoyiannis, 2012; Sikorska et al., 2012, 2015, Smith et al. 2015). Alternative methods condition model errors on input and/or output variables (Del Guidice et al., 2013; McInerney et al., 2017), or link their parameters (i.e., of the error distributions) to flow conditions (Schaefli et al., 2007), or introduce their time-dependent description (Pianosi & Raso 2012). Alternatively, several methodologically simpler likelihood-free approaches are frequently applied that rely on metaheuristic search algorithms (Piotrowski et al. 2017) such as Genetic Algorithm, Particle Swarm Optimization, where model deficiencies are accounted for via using multiple generated parameter sets instead of making assumptions on model errors.

The second group of methods relies on data-driven models (DDMs) that establish statistical links between target and input variables using a training dataset but do so without including any classical hydrological model (Bowden et al., 2005, Solomatine & Ostfeld 2008). Thus, DDMs are very powerful for modelling complex relationships between input and output variables or for gaining insights on governing (physical) processes (Tongal & Booij, 2018)

without the need to explicitly consider the physical laws governing such processes. Hence, they are constantly increasing in popularity within hydrology and have been applied for many different hydrological applications: precipitation-runoff modelling (Rajurkar et al., 2004; Shortridge et al., 2016; Tongal & Booij, 2018), streamflow forecasting (Solomatine & Xue 2004; Boucher et al., 2011; Papacharalampous & Tyralis, 2018; Ni et al., 2020; Tyralis et al., 2020), groundwater level forecasting (Suryanarayana et al., 2014; Rahman et al., 2020), water quality modeling (Fatehi et al., 2015; Bhagat et al., 2021), and many others. DDMs require however the pre-selection of input variables as potential predictors, which greatly impacts their accuracy (Galelli et al., 2014). In addition, since hydrological variables often display auto- and cross-correlation it is important to consider time-lagged versions of the input variables (e.g., precipitation, air temperature) when simulating a target hydrological variable (e.g., streamflow) (Gauch et al., 2021), and to be able to identify the maximum lag time (memory length). This maximum time lag defines a lag time after which the impact of the input variable on the target variable is marginal (Bowden et al., 2005). A correct pre-selection of input variables enables the exclusion of redundant and/or irrelevant input variables, thereby reducing the complexity and increasing the interpretability of the resulting DDM (Quilty et al., 2016).

Most previous data-driven approaches for streamflow simulation have used DDMs to explicitly simulate streamflow and in this way they provide an alternative to a HM. Yet, only very few works have attempted to characterize or directly simulate residuals of a HM with a data-driven approach. Some examples include Montanari & Koutsoyiannis (2012), Sikorska et al. (2015), Wani et al. (2017), and Ehlers et al. (2019), who simulate model errors via resampling from a query dataset based on streamflow properties (and other hydro-meteorological data in the case of Ehler et al., 2019). However, each of these studies relied on a specific statistical method for resampling model errors (the meta-Gaussian approach in Montanari & Koutsoyiannis (2012) and k nearest neighbours in the others) and, therefore, their approaches are not generalized to any DDM (as is the case in this study).

Approaches for coupling HMs with DDMs into a hybrid model are very rare in application to streamflow modelling (Wang et al. 2021). For example, Tongal and Booij (2018) decomposed streamflow into base and surface flows and used HM simulations at the previous time step along with meteorological variables as input to a DDM to simulate streamflow at the current time step, indirectly linking model residuals to flow conditions. Senent-Aparicio et al.

(2018) have proposed an indirect coupled approach, in which streamflow simulations from a DDM are corrected using additional information on maximum mean daily flow that is taken from a HM. Yang et al. (2020a) have proposed a different indirect coupled approach that used a HM to simulate pseudo-observed data for training a DDM to perform streamflow simulation. Hybrid streamflow simulation models (whereby a DDM is used to model HM residuals), have only been explored by Wu et al. (2019) and Konapala et al. (2020). Both studies used DDMs for simulating the residuals of HMs but they both lack consideration of the uncertainty in estimating HM parameters. Thus, the authors in these two studies did not consider ensemble streamflow simulations. In addition, Wu et al. (2019) required the transformation of HM residuals (prior to modelling via DDMs), and the determination of stationary time windows (as identified by the Hilbert Huang Transform) in order to simulate the HM residuals, and assumed that the simulated HM residuals follow a Student's t-distribution (to permit the construction of uncertainty intervals). More recently, Wang et al. (2021) proposed a hybrid model that uses the output of the Xinanjiang HM along with wavelet-decomposed sub-series of previous streamflow observations as input to Random Forests for simulating a single realization of streamflow without any uncertainty considered. Althoff et al. (2021) proposed an alternative hybrid model that consisted of a simplified version of a HM (soil moisture component) and a DDM to simulate streamflow. However, they also provide only a single realization of the streamflow without considering any sources of uncertainty.

As can be seen from the above literature review, despite several recently proposed hybrid modelling approaches, none of the above mentioned studies developed a fully coupled framework for an *ensemble-based streamflow simulation* where a DDM is used to simulate an ensemble of residuals stemming from a HM, conditioned on input variables (e.g., precipitation, air temperature). Thus, in this study, a novel conceptual-data-driven modelling framework is developed that pairs an ensemble of conceptual deterministic HMs with an ensemble of DDMs-(that simulate the HM residuals) for improved ensemble streamflow simulation. This novel framework is called an *ensemble-based conceptual-data-driven approach* (CDDA). While a HM simulates the precipitation-streamflow process at the catchment scale, respecting, to an extent, the physical hydrological processes, a DDM, added 'on-top' of the HM, mimics the HM residuals enabling the streamflow simulations to be improved. Thus, the CDDA combines the advantages of a HM (that seeks to respect the physical relationships between hydrological

processes) with the capabilities of the DDM (that can model complex (nonlinear) relationships between input-target variables) enabling an effectual estimation of the correlated residuals stemming from the HM. Evidently, the resulting output of the CDDA framework is an ensemble of streamflow simulations instead of a single streamflow realization. Conceptually, the HM-generated ensemble passes information about the observed and simulated streamflow (gained through the identification of the HM parameters, i.e., calibrated using observed data) to the DDM, which is relied upon to extract additional information not captured by the HM, to simulate the HM residuals and improve the overall ensemble streamflow simulation. Thus, the CDDA can overcome certain issues of the above mentioned hybrid models since it provides an ensemble of HM residuals (accounting for HM parameter uncertainty), and does not require any transformation of HM residuals, relaxing all assumptions on the residual distribution. These last two strengths of the proposed CDDA approach also make it more attractive than the Bayesian method (described above), that requires specification of a likelihood function and the distribution of the model residuals. Thus, the novel CDDA approach can be seen as a less-restrictive and more flexible data-driven method for simulating HM residuals and generating an ensemble of streamflow simulations.

With regards to the above, the focus of this paper is primarily on developing the CDDA framework, that enables, for any case study where data required by the HM is available, the identification of the best DDM to simulate HM residuals and the selection of the most useful input variables to use in the DDM. Since any DDM can be used within the CDDA, eight different DDMs, that have either been explored in detail or shown to be promising in recent studies in the hydrological modelling literature, are explored: Multiple Linear Regression (MLR), k Nearest Neighbours Regression (KNN), Second-Order Volterra Series Model (SOV), Artificial Neural Networks (ANN), and two variants of eXtreme Gradient Boosting (XGB) and Random Forests (RF). These models are developed by considering time-lagged copies of observed streamflow, precipitation, and air temperature as potential predictors and are coupled with a bucket type HM, Hydrologiska Byråns Vattenbalansavdelning or HBV (Bergström & Forsman, 1973), and tested using three Swiss catchments. Furthermore, several input variable selection (IVS) methods are considered for selecting the most important candidate input variables to use in the DDM. Thus, the major objective of this study is to introduce the ensemble-based CDDA framework and evaluate its performance against a standalone (ensemble-

based) HM model. In addition, the proposed CDDA enables one to answer questions specific to individual case studies, such as: (A) Which DDMs are most suitable for simulating the residuals of the HM within CDDA? (B) Which input variables are the most important to consider when simulating HM residuals via the explored DDMs? The framework developed in this study is timely, as coupled HM-DDM approaches are only beginning to consider uncertainty in resulting streamflow simulations (e.g., Papacharalampous et al., 2019a; Tyralis et al., 2019a; Boucher et al., 2020; Teweldebrhan et al., 2020), although no such framework exists that uses DDMs to simulate the residuals from an ensemble of HM streamflow simulations.

The remainder of this paper is organized as follows: Section 2 describes the methods by introducing the theoretical CDDA framework and providing details on candidate DDMs selected for simulating the HM residuals; Section 3 includes the experimental settings, an overview of the case study, and the HM and DDM development details; Section 4 highlights the main results; Section 5 discusses the significance of the results, describes the current limitations of the developed approach, and suggests future research avenues; and Section 6 concludes the paper.

2 Methods

2.1 Overview of the Conceptual-Data-Driven Approach (CDDA)

A single data-driven model is developed that mimics residuals of a hydrological model and is linked to the input (explanatory) meteorological variables that are usually precipitation and air temperature. This DDM is attached on top of a single simulation (e.g., streamflow) from the HM resulting from its parameter set. Thus, the target (response) variable is the residual between the observed streamflow and the HM-simulated streamflow at the time step t-0. In case of multiple parameter sets of the HM (i.e., an ensemble of HM streamflow simulations), the DDM is attached to each HM simulation (i.e., there is a single DDM trained for each set of residuals). This latter models' setup, i.e., an ensemble HM+DDM, is called an *ensemble-based conceptual-data-driven approach*, see Figure 1 for an overview.

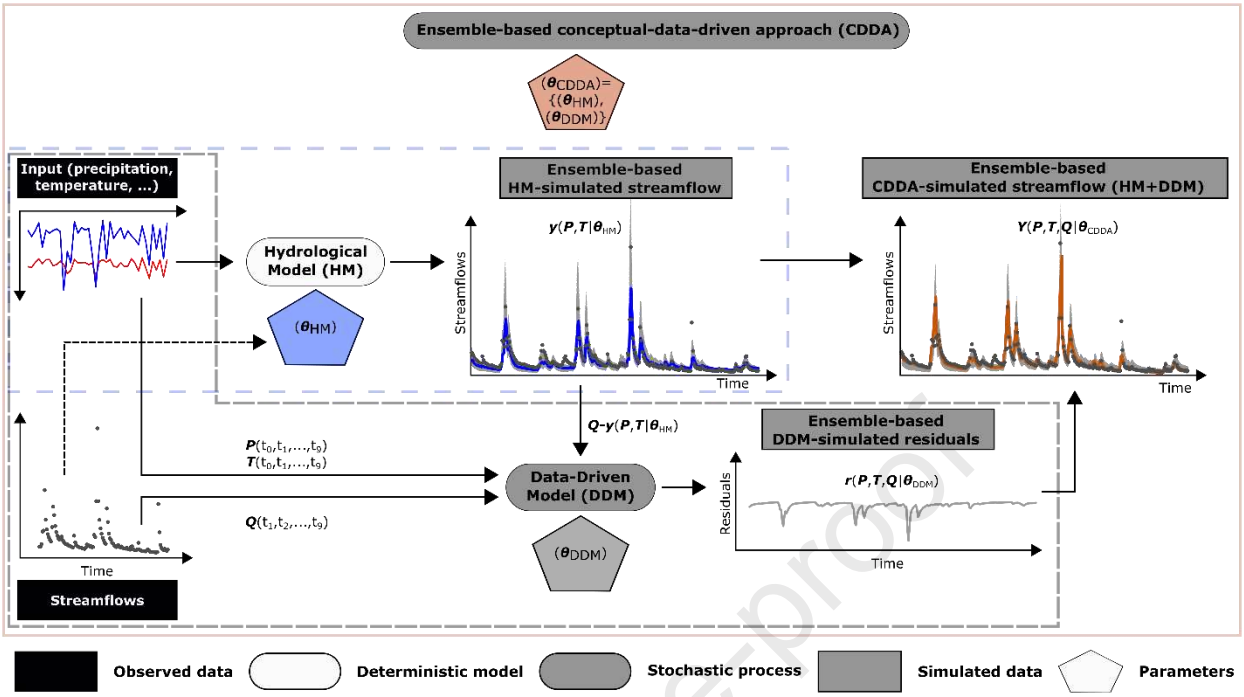


Figure 1. Schema of the developed ensemble-based conceptual-data-driven approach (CDDA) (see Section 2.2 for additional information on notation).

The residuals of the HM are modelled using different DDMs taking the meteorological variables and observed streamflow at previous days as an input (see Section 2.3 for details). Different input combinations are tested (Section 2.4) to explore which set of inputs best simulate the residuals of the HM. Note that the observed streamflow is assumed to be available (at a regular time interval) at a given site. In case observed streamflow is unavailable, both the HM and DDM cannot be calibrated and other indirect approaches would be required (see Section 5.2).

To evaluate the usefulness of the CDDA, the ensemble of deterministic outputs from the hydrological model without any DDM is used as the benchmark (Section 2.6). The eight different variants of the CDDA are proposed and compared against each other, and the benchmark, to determine the best DDM and identify useful input combinations.

2.2. Ensemble-based CDDA

The streamflow simulation for the ensemble member Y_i of the CDDA output is the sum of the simulation of the streamflow with the HM model (y_i) and the simulation of its residual using the DDM model (r_i):

$$Y_i = y_i + r_i$$

$$Y_i(\boldsymbol{P}, \boldsymbol{T}, \boldsymbol{Q} | \boldsymbol{\theta}_{CDDA}) = y_i(\boldsymbol{P}, \boldsymbol{T} | \boldsymbol{\theta}_{HM}) + r_i(\boldsymbol{P}, \boldsymbol{T}, \boldsymbol{Q} | \boldsymbol{\theta}_{DDM}) \quad (1)$$

where $\boldsymbol{P}$, $\boldsymbol{T}$, and $\boldsymbol{Q}$ are observed precipitation, air temperature, and streamflow, $\boldsymbol{\theta}_{CDDA}$ is a parameter set of the CDDA, $\boldsymbol{\theta}_{HM}$ and $\boldsymbol{\theta}_{DDM}$ are the parameter sets of the HM and DDM models, respectively, with $\boldsymbol{\theta}_{CDDA} = \{\boldsymbol{\theta}_{HM}, \boldsymbol{\theta}_{DDM}\}$.

Note that each ensemble member is associated with a single $\boldsymbol{\theta}_{CDDA}$. By applying Eq. (1) to each ensemble member, an ensemble of streamflow simulations CDDA can be obtained through CDDA. Finally, while the HM requires $\boldsymbol{P}$ and $\boldsymbol{T}$ as input to simulate streamflow for a given day, the DDM can additionally use $\boldsymbol{P}$, $\boldsymbol{T}$, and $\boldsymbol{Q}$ from previous days (i.e., time-lagged copies of these variables) in order to improve predictive performance. This idea is discussed in more detail in Section 2.4.

The HM is run at a daily time step, as is the DDM. Depending on the chosen HM model structure, other input variables may be also required for the HM model (e.g. potential evapotranspiration) and thus, could also be explored within the DDM. Similarly, using a different time step for the HM (e.g., hourly) may require the DDM to be conditioned on input variables of a different resolution and different lag time length.

2.3 Data-Driven Models (DDM)

In this study, six different DDM (and eight variants, in total) were explored for simulating the residuals of the HM. Since this section is intended to provide only a brief explanation of each DDM, references to appropriate literature are included for the reader interested in a more detailed treatment of the various methods. However, due to the ubiquity of Multiple Linear Regression (MLR) (which is one of the adopted DDMs), no details are provided for this model.

2.3.1 K Nearest Neighbours Regression (KNN)

K Nearest Neighbours regression (KNN), introduced first by Fix and Hodges (1951), is one of the simplest nonlinear methods to apply (Altman 1992, Mitchell 1998, Liu et al. 2004, Hastie et al. 2009). KNN generates predictions for a 'query' vector (i.e., an input variable vector) by applying two simple steps: 1) searching for a set of K neighbours (i.e., input variable vectors) in a training dataset that are closest to the query according to a predefined distance metric (e.g., Euclidean distance) and 2) taking the average of the target (response) variable associated with each of the K closest neighbors. Thus, KNN is a local regression technique.

Despite its simplicity, one of the main drawbacks of KNN is that it cannot extrapolate beyond the range of the target in the training set. Nonetheless, KNN has a rich history in hydrology (Karlsson & Yakowitz, 1987) and is still frequently used for simulating and predicting streamflow (Lee et al., 2017; Ebrahimi et al., 2020), among other hydrological applications (Sun & Tevor, 2017; Jiang et al., 2020). For more information on KNN, the interested reader may refer to Chapter 2.3 of Hastie et al. (2009).

2.3.2 Second-Order Volterra Series Model (SOV)

Similar to KNN, the Second-Order Volterra series model (SOV) is another simple nonlinear regression technique. SOV was initially proposed for measuring the (non)linearity of hydrologic systems (Amorocho, 1963) and subsequently applied for rainfall-runoff simulation (Amorocho & Brandstetter, 1971; Diskin et al., 1984) and more recently for forecasting streamflow (Maheswaran & Khosa, 2012), groundwater levels (Maheswaran & Khosa, 2013), urban water demand (Quilty & Adamowski, 2018) and soil moisture (Prasad et al., 2018) as well as downscaling climatic variables (Seghal et al., 2016, 2018; Lakhanpal, 2017).

SOV, is essentially a polynomial regression where the design matrix is created by considering all zero-, first-, and second-order interactions amongst the input variables. The coefficients attached to the zero-order term is the bias and the remaining coefficients (attached to first- and second-order terms) are considered kernel coefficients (Wu & Kareem, 2014). After they have been estimated, the coefficients can be used to generate predictions of the target variable for a given input variable vector.

The reader may consult the Supplemental Material of Quilty and Adamowski (2020) for the mathematical formulation of the SOV adopted in this study.

2.3.3 Artificial Neural Networks (ANN)

Artificial Neural Networks have been applied to streamflow prediction since the 1990's (Abrahart et al., 2012) and represent one of the most popular data-driven models used for hydrological prediction (Fahimi et al., 2017). ANN generates predictions by passing a vector of input variables through a network comprising of a series of weights that connect to neurons, which contain a bias and (potentially, a nonlinear) activation function, that transforms the weighted inputs and directs them to the next layer of the network - the output of the network is the sum of weighted and (nonlinearly) transformed inputs. The reason why ANN is so powerful is due to its universal approximation capabilities, which allow it to approximate any nonlinear function (i.e., mapping from input variables to the target) provided nonlinear infinitely differentiable activation functions are used in the hidden layer and enough training iterations (epochs) as well as model parameters (weights and biases) are considered (Hornik et al., 1989).

The type of ANN used in this study is a feed-forward backpropagation (FFBP) network with a single hidden layer (along with the usual input and output layers, representing the input variables and target, respectively). The FFBP ANN operates by initializing all network parameters (i.e., weights and biases) to small random values and iteratively updates these parameters by back propagating the error signal through the network for a fixed number of epochs or until sequential evaluations of the error function do not appreciably decrease with respect to a predefined tolerance. After it has been trained, ANN generates predictions by passing input variable vectors through the network, where the resulting output is the sum of weighted and (nonlinearly) transformed inputs.

For additional information on ANN, the reader can refer to Chapter 5 of Ripley (1996) or many of the references included in review articles in the domains of statistics (Ching & Titterington, 1994) or hydrology (Abrahart et al., 2012; Wu et al., 2014).

2.3.4 Random Forests (RF)

Random Forests, introduced by Breiman (2001), are a class of (ensemble) decision trees that include random sampling of both training instances (often referred to as bagging) and input variables, reducing predictive variance (without increase the bias) of the model (Breiman, 1996) while also providing robust performance in the presence of noisy input variables (Biau, 2012). A very useful feature of RF is that it implicitly provides an estimate of input variable importance

(Grömping, 2009), which can be used for input variable selection (Genuer et al., 2010), see for example Tyralis & Papacharalampous (2017). RF generates predictions by taking the average across all predictions produced by each ensemble member.

Despite its widespread success in numerous domains, such as genomics (Chen & Ishwaran, 2012), remote sensing (Belgiu & Drăgu, 2016), RF is only beginning to grow in popularity within hydrology (Tyralis et al., 2019b). Although RF is still relatively new to the hydrology domain, it has been used for streamflow simulation (Shortride et al., 2016), forecasting (Papacharalampous & Tyralis, 2018), and reconstruction (Li et al., 2019), groundwater level forecasting (Rahman et al., 2020), and downscaling extreme rainfall (Pham et al., 2019), among other applications.

The interested reader may refer to Chapter 15 of Hastie et al. (2009) for details on RF as well as Fawagreh et al. (2014) for a review of different RF variants and applications.

2.3.5 eXtreme Gradient Boosting (XGB)

The eXtreme Gradient Boosting model is a very recent method that combines decision trees and boosting (Chen & Guestrin, 2016). While bagging (as used in RF) operates by developing an ensemble of independent models on each resampled dataset, boosting instead builds an ensemble of models where each member is built on top of the residuals generated by the previous member, with the objective to reduce the errors made in the previous iteration. In other words, bagging reduces the variance of the predictors, while boosting reduces bias (Vanschoren et al., 2012); both addressing opposite aspects of the bias-variance trade-off and using different strategies for 'ensembling' decision trees (Mehta et al., 2019). While recent research has shown that combining bagging and boosting can lead to better performing ensemble decision trees (Ghosal & Hooker, 2020), such methods are not within the scope of this study.

XGB is an improved version of the gradient boosting machine that is more computationally efficient and less prone to overfitting due to L2-norm regularization (Chen & Guestrin, 2016). Similar to RF, XGB inherently estimates the importance of input variables and can be used for input variable selection (e.g., Chen et al., 2020; Rahman et al., 2020; Zhang et al., 2020).

XGB has only recently been considered within the hydrological domain but has shown promise for streamflow simulation (Hadi et al., 2019; Gauch et al., 2021) and forecasting (Ni et al., 2020; Tyralis et al., 2020), predicting daily reference evapotranspiration (Fan et al., 2018), water quality index prediction (Abba et al., 2020), water table depth forecasting (Brédy et al., 2020), and imputing missing sub-hourly precipitation records (Chivers et al., 2020), among other applications.

Additional details on the theory and main innovations behind XGB can be found in Chen & Guestrin (2016).

2.4 Input Variable Selection (IVS)

Since input variable selection is an important step in the development of any DDM (Galelli et al., 2014), for each of the six DDMs (MLR, KNN, SOV, ANN, RF, and XGB), IVS was carried out to identify the input variables that may be useful for simulating the HM residuals. Different IVS approaches were considered for the DDM. The (linear) partial correlation input selection (PCIS) algorithm (May et al., 2008) was paired with MLR to serve as a fully linear benchmark. The Edgeworth approximations-based approach (EA) was used to carry out CMI-based IVS, which is a nonlinear analogue to PCIS (Quilty et al., 2016). The EA approach was coupled with KNN, SOV, and ANN. Since both RF and XGB implicitly perform IVS, an 'external' (model-free) IVS method (e.g., PCIS, EA) was not necessary. The PCIS method was selected as it is one of the most popular linear IVS methods (Galleli et al., 2014) while the EA method was selected as it has been shown to provide similar IVS accuracy as competing CMI-based methods (e.g., based on kernel density estimation, k nearest neighbors) at a fraction of the computational 'cost' (Quilty et al., 2016). A short description of the different IVS methods is provided in the sub-sections below.

2.4.1 Partial Correlation Input Selection (PCIS)

PCIS is an iterative greedy IVS method whereby candidate input variables (i.e., time lagged copies of observed streamflow, precipitation, and air temperature) are selected one at a time based on their partial correlation with the target variable (i.e., HM residuals at t-0), conditioned on previously selected inputs. At each iteration, the candidate input variable with the highest partial correlation (i.e., the best candidate input variable) is selected and a predefined IVS stopping condition is checked. The version of PCIS adopted in this study uses the Bayesian

Information Criterion (BIC) as a stopping condition to halt the IVS procedure (see Galleli et al., 2014). At each iteration of the IVS procedure, an MLR model is built that predicts the target variable using all previously selected inputs and the best candidate input variable; afterwards, the BIC is measured. If the BIC increases for the current iteration compared to the previous one, the IVS procedure is halted and all input variables selected *before* the current iteration are returned. Otherwise, the IVS procedure continues. The variable importance measure for PCIS is the partial correlation coefficient (PC), which ranges between -1 and 1 (with 0 representing independence while -1 and 1 represent perfect correlation).

Additional details on PCIS can be found in May et al. (2008) and Galleli et al. (2014).

2.4.2 Edgeworth Approximations-based Conditional Mutual Information (EA)

The EA approach to CMI-based IVS was introduced and discussed in detail in Quilty et al. (2016); thus, only the essential features of this method are described here. Similar to PCIS, EA is an iterative greedy IVS method. However, in EA, the CMI is estimated between candidate input variables and the target, conditioned on previously selected inputs, instead of partial correlation. The stopping condition used for EA is based on a modified version of the tolerance-based approach from Vlachos & Kugiumtzis (2010) where at each iteration the CMI between the best candidate input variable and the target variable, conditioned on all previously selected inputs is compared to the mutual information between the target variable and all previously selected input variables, including the best candidate input variable. If the ratio between these two quantities drops below the tolerance (ranging between 0 and 1), the IVS procedure halts and all input variables selected *before* the current iteration are returned. By increasing the tolerance, the number of selected input variables can be decreased. The only difference between the EA method adopted here and in Quilty et al. (2016) is in the choice of stopping criterion. It was found during earlier experimentation that the tolerance-based method provided greater control over the IVS process, facilitating an improved balance between computational run-time and model accuracy (compared to the method used in Quilty et al., 2016). Trial-and-error led to the selection of 0.05 as a suitable threshold to balance these two objectives. Typically, for CMI-based IVS adopting the tolerance threshold stopping criterion, values of 0.01, 0.03, 0.05, and 0.15 are common (Vlachos & Kugiumtzis, 2010; Tsimpiris et al., 2012). (Note: that these tolerance values are obtained by subtracting the tolerance values mentioned in the above studies

from 1; the reason for this is due to our modified formulation of the stopping criterion presented in Vlachos & Kugiumtzis, 2010).

The variable importance measure for the EA method is the partial informational correlation (PIC), which is a nonlinearly scaled version of CMI (Sharma & Mehrotra, 2014) such that the PIC ranges between 0 and 1 (with 0 representing independence between variables and 1 indicating perfect correlation).

2.4.3 Variable Importance (Decision Tree-based) Methods

The decision tree-based methods (RF and XGB) generate internal measures of variable importance that are useful in identifying the most important inputs to the model (Wang et al., 2018; Pathy et al., 2020). In contrast to the PCIS and EA methods, which are considered model-free approaches, RF and XGB are model-based IVS approaches. While the input variable importance measures generated by RF and XGB can also be used for IVS in other DDM (see for instance, Chen et al., 2018; Hadi et al., 2019; Prasad et al., 2019; Chen et al., 2020; Tyralis et al., 2020; Bhagat et al., 2021), this study only uses the variable importance measures to identify the most useful variables specific to these models.

For RF, the variable importance score is the total decrease in node impurity that occurs by splitting on a particular input variable and averaged over all decision trees, measured by the sum of squared errors (Li et al., 2017; Wright & Ziegler, 2017). The variable importance scores generated by RF were normalized by dividing each input variable's score by the maximum variable score (see Eq. (1) in Deng & Runger (2013)), resulting in the normalized variable importance (NVI).

The variable importance (VI) score adopted for XGB is the fractional contribution of each input variable to the model prediction, averaged across all trees, also known as 'gain' (Li et al., 2017), where the model's predictive performance is measured by the sum of squared errors loss function (Chen et al., 2019). For both RF and XGB, higher importance scores represent variables that are more important than the others (Li et al., 2017).

2.5 Benchmark Method

As discussed in Section 2.1, the CDDA can generate an ensemble of streamflow simulations via Eq. (1). In order to evaluate the added value of using the CDDA it must be

compared against a benchmark. In this case, the benchmark is simply the ensemble streamflow simulations generated by the HM model for each $\boldsymbol{\theta}_{HM}$ (i.e., without any model for simulating the HM residuals).

2.6 Performance Metrics

In order to compare the performance of CDDA against the benchmark HM, as well as to identify the best DDM to use within the CDDA, a number of statistical (performance) metrics were used. The performance metrics were divided into two classes: ensemble and deterministic metrics. The ensemble metrics make use of all ensemble members when evaluating the simulations' performance while the deterministic metrics are computed using the mean ensemble member, i.e., the mean simulation at each simulation time step. Since the metrics adopted in this study are well-known in the hydrology community, the formulae used to calculate these scores are not provided although the cited sources include the necessary information to permit their calculation.

The ensemble metrics consist of the mean continuous ranked probability score (CRPS) (Bröcker, 2012), the mean interval score (MIS) (Gneiting & Raftery, 2007), and the average width (AW) (Xiong et al., 2009). Both the MIS and AW require the specification of a confidence level, which is taken to be 0.05 in this study. For a particular confidence level, upper and lower uncertainty intervals can be computed, either by assuming the ensemble members follow a particular distribution (e.g., Gaussian) or by empirical means, such as estimating the quantiles associated with the confidence level. The latter approach is followed in this study; thus, the upper and lower uncertainty intervals are estimated at the 0.975 and 0.025 quantiles, respectively, and they together define the 95% uncertainty intervals. The CRPS was calculated using the scoringRules R package (Jordan et al., 2019) while the MIS and AW were calculated using custom R functions.

The CRPS is a useful metric that evaluates a simulation's reliability and sharpness, reducing to the mean absolute error (MAE) for deterministic simulations, permitting the comparison between ensemble and deterministic forecast quality (Boucher et al., 2011). The MIS also evaluates a simulation's reliability and sharpness, but additionally includes a penalty for simulations that fall outside the upper and lower uncertainty intervals (Papacharalampous et al., 2020). The AW solely measures the simulation's sharpness (i.e., the narrowness of its

uncertainty intervals). In general, a high quality ensemble simulation should have CRPS, MIS, and AW as low as possible. However, preference in this study is given to simulations with lower CRPS and MIS scores since a low AW score is not highly informative if the simulations are unreliable. Finally, since the CDDA seeks to improve the predictive performance of the ensemble HM simulations (which themselves only consider parametric uncertainty) through DDM, the term uncertainty intervals (as adopted here) is more closely related to confidence intervals (CIs) than prediction intervals (thus, CIs are referred to when the model results are discussed). Typically, prediction intervals take into account the uncertainty in the model output, which is not done here since we use the DDM to predict the expectation of the individual HM errors, not their distribution.

The deterministic metrics adopted for evaluating the ensemble models include the mean absolute error (MAE), root mean square error (RMSE), Nash Sutcliffe Efficiency (NSE), and the Kling-Gupta Efficiency (KGE). The advantages of these metrics are described in detail within the Supplemental Material of Papacharalampous et al. (2019b). The deterministic metrics were calculated using the hydroGOF R package (Zambrano-Bigiarini, 2017).

3 Experimental Settings

3.1 Study Catchments

Three Swiss mid-size mid-altitude catchments were chosen for this study (Figure 2). All three catchments are with an insignificant areal glacier percentage (<5%) and without any significant human direct impacts documented in the observation period (1981-2015). According to Sikorska-Senoner & Seibert (2020), the catchments represent two different types of dominant flood processes, i.e. rainfall-driven (Dünnern) and a mixed contribution of rainfall and snowmelt floods (Kleine Emme and Muota), see also Table 1.

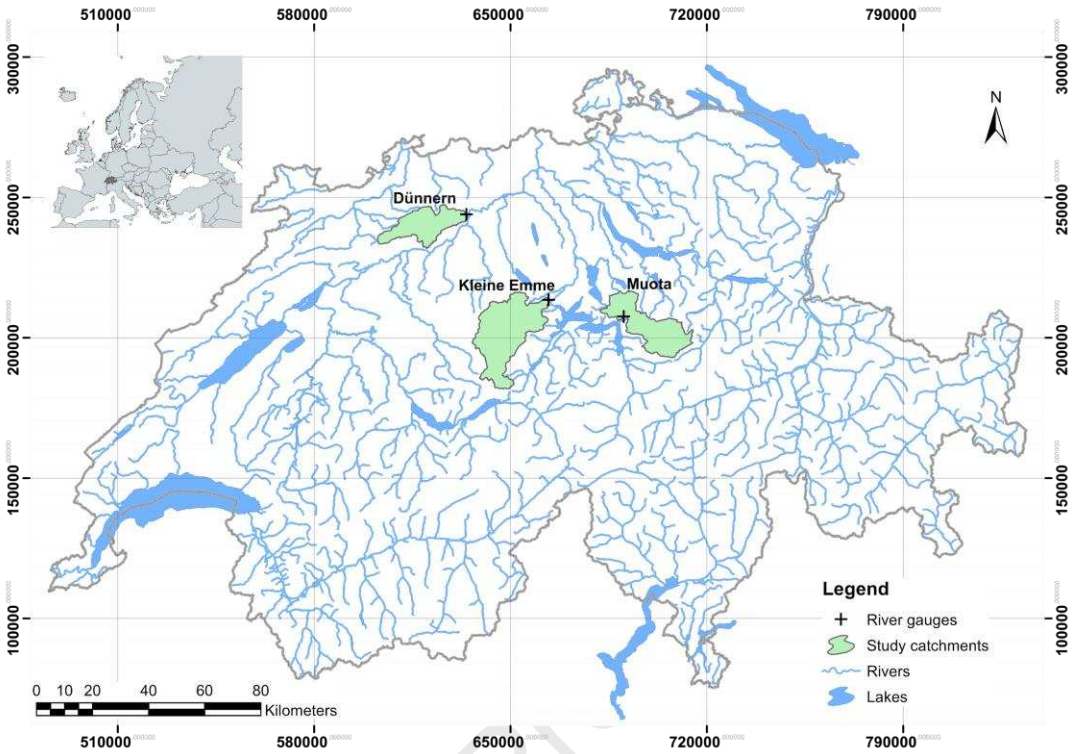


Figure 2. Location of 3 study catchments, Swiss coordinate system.

Table 1. Characteristics of the study catchments by an increasing mean elevation.

ID	Catchment	Gauging station	Area (km^2)	Mean altitude (masl)	Dominant flood type
C1	Dünnern	Olten	234	711	rainfall
C2	Kleine Emme	Emmen	478	1054	rainfall & snowmelt
C3	Muota	Ingenbohl	317	1363	rainfall & snowmelt

3.2 Observed Data

The available observed data consist of continuous records of precipitation depth (mm/day), minimal, mean and maximal daily temperature (°C), daily evaporation rates (mm/day)

and streamflow at catchment outlets (mm/day). All variables were available for the period 1981-2014 at a daily resolution. The meteorological data were made available from the MeteoSwiss and the hydrological observations from the Swiss Federal Office for the Environment (FOEN). All variables were averaged to mean catchment values using the Thiessen polygons' method.

3.3 Hydrological Model

In this study, a conceptual HBV model, in particular, HBV light (Seibert & Vis, 2012), was used for streamflow simulations. This bucket-type model consists of four major routines: i) precipitation excess and snowmelt, ii) soil moisture, iii) groundwater and runoff-streamflow response, and iv) routing in the stream. The snow component is important for catchments with significant snow processes, i.e., for two out of three studied here. HBV has been frequently applied to rainfall- and snow-dominated catchments (e.g., Breinl, 2016; Griessinger et al., 2016; Sikorska & Seibert, 2018; Westerberg et al., 2020).

The HBV model used in this study has 15 parameters and is run at a daily time step. Model inputs are precipitation and air temperature time series and long term averaged values of daily evaporation and air temperature. The model output is a single (deterministic) realization of continuous streamflow time series at the catchment outlet. Due to its simplicity, HBV is used as an example in this study but can be easily interchanged with another HM without any difficulties.

3.4 Calibration of the Hydrological Model

The HBV model was calibrated via the Genetic Algorithm and Powell (GAP) optimization method (Seibert, 2000) using the Kling-Gupta efficiency (KGE) as the objective function (Gupta et al., 2009). Calibration involved 1000 independent runs with randomly selected initial values for the 15 parameters, resulting in 1000 optimized and equally plausible parameter sets representative of model parametric uncertainty. This uncertainty mainly results from parameter equifinality (Beven & Freer, 2001) (and to a lesser degree the randomization involved in initializing the parameter sets). Running multiple independent model calibration runs with randomly chosen initial values has been proposed by Sikorska-Senoner et al. (2020) as a heuristic approach to deal with the parameter equifinality problem. Such an optimization approach should ensure that the parameter space is fully explored and should minimize the possibility of being trapped in the same local optimum during different calibration runs. Such an

approach is prioritized over the likelihood-based methods, since it does not require making any assumption about the model errors, which are simulated with a dedicated DDM.

The HBV model was calibrated using years 1985-2004 (with a preceding 4-year warm-up period of 1981-1984) and validated using years 2005-2014 at a daily time step in the three study catchments. The KGE score achieved in the validation period (Table 2) for the mean ensemble varied from 0.80 in the Dünnern catchment to 0.87 in the Muota and Kleine Emme catchments. These 1000 optimized parameter sets are next used within HBV to generate an ensemble of streamflow simulations and subtracted from the observed streamflow to generate an ensemble of HM residuals. Afterwards, the DDMs are calibrated using each of the 1000 HM residuals and the inputs determined through IVS.

3.5 Input Variable Selection and Calibration of Data-Driven Models

Prior to the calibration of DDMs, input variable selection was required. Input variables for the DDM included observed streamflow, precipitation and air temperature, at the current and/or previous nine days. For streamflow, inputs included observations from the nine days preceding the simulation day (t-1, ..., t-9). For precipitation and air temperature, inputs included observations from the day of the simulation (t-0) as well as the previous nine days (t-0, ..., t-9). In total, 29 different input variables were considered as potential predictors of the HM residuals at t-0.

The maximum time lag (D) for each input variable was determined using the conditional mutual information (CMI) (Brown et al., 2012) between the HM residuals at t-0 and each explanatory variable from t-0 to t-D (with the exception of streamflow, which considered time lags t-1 to t-D) by locating the lag at which the CMI reached a local minimum. The goal was to identify a sufficient number of time lags to accurately simulate the HM residuals while also attempting to keep the input variable set of a reasonable size. Given that precipitation has the largest effect on modifying the streamflow (Müftüoğlu, 1991), it was given a higher priority in identifying the maximum time lag. This approach resulted in a maximum time lag of D=9. Section 2.4.2 outlines the method used to estimate CMI.

The different DDMs were calibrated (trained) using the target (residuals of the HM-simulated streamflow) and input variables (i.e., time-lagged versions of the observed precipitation, air temperature, streamflow) for the same calibration period as the HBV model, i.e.

years 1985-2004. The remaining data (years 2005-2014) was used for validating the DDMs. For consistency, the DDMs used the same calibration and validation periods as the HM. In the sub-sections below, the different DDM parameters and hyper-parameters are described along with the strategy used to train the various DDMs. All DDMs were developed on an Intel(R) Core(TM) i7-8750H CPU @2.20 GHz laptop with 32.0 GB RAM. The ANN, RF, and XGB models were run in parallel on 10 CPU cores.

3.5.1 K Nearest Neighbours Regression (KNN)

As noted in Section 2.3.1, the KNN model does not require any explicit training strategy since model predictions are generated by computing the distance (here, the Euclidean distance) of a given input variable vector (e.g., from the validation set) with those from the training data, locating its K nearest neighbours, and taking the mean of the targets associated with each of the K neighbours. While KNN does not require training, it requires the selection of the hyper-parameter, K, with lower K values leading to predictions with high variance and low bias and vice versa for high K values (Hastie et al., 2009). Previous research by one of the authors found that a wide variety of K values (5, 10, 20, 30, 50, and 100) lead to similar results when sampling conditional model errors for streamflow simulation (Sikorska et al., 2015). Thus, this study used K=5, seeking to strike a balance in the bias-variance tradeoff related to the selection of K. Other common choices for K include $\sqrt{N}$, where N is the number of samples in the training set (Lall & Sharma, 1996).

To ensure input variables with higher ranges are given equal weight (or importance) as input variables with smaller ranges, all inputs are individually normalized (i.e., scaled between 0 and 1, using the maximum of minimum of each variable over the training set), prior to searching for nearest neighbours. Using normalized inputs in KNN has been shown to significantly improve model performance (compared to using unnormalized inputs) (Piryonesi & El-Diraby, 2020) and in earlier experiments was also found to have the same effects for the catchments under study. The FNN package in R (Beygelzimer et al., 2019) was used for developing the KNN models, which uses the fast k nearest neighbours method (Beygelzimer et al., 2006) and the kd-tree approach (Friedman et al., 1977) when searching for nearest neighbours.

3.5.2 Multiple Linear Regression (MLR) and Second-Order Volterra Series Model (SOV)

The parameters in MLR are the slope and bias coefficients associated with the design matrix (input variables). While the parameters in SOV are the bias (zero-order) and kernel coefficients associated with the first- and second-order interactions amongst the input variables. The parameters in MLR and SOV were solved via ordinary least squares (Wu & Kareem, 2014).

Note that there are no tunable hyper-parameters for the MLR and SOV models. The MLR and SOV models were developed using custom functions in R.

3.5.3 Artificial Neural Networks (ANN)

The parameters in ANN include the input, hidden, and output layer weights as well as the hidden and output layer biases. The particular application of ANN used in this study is based on the avNNet function in the caret R package (Kuhn, 2019), which makes use of the nnet R package (Venables & Ripley, 2002). This implementation of ANN individually trains several networks with different randomly initialized parameters via the Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm (Fletcher, 1987) and averages their predictions. The hyper-parameters include the number of networks (set by default to 5), the number of training epochs (or iterations, set by default to 100), the number of hidden nodes (since only a single hidden layer is used), and the decay rate.

A grid search over the decay rate (1e-3, 0.01, 0.1) and number of hidden neurons ($1{:}\,2N_I + 1$; where N_I is the number of inputs in the ANN) (Hecht-Nielsen, 1989; Fatehi et al., 2015) was carried out using 5-fold cross-validation and the RMSE as the objective function, to determine optimal values for these hyper-parameters. The ANN models adopted linear activation functions in the output layer and sigmoid (logistic) activation functions in the hidden layer. All model inputs were normalized before training the ANN (i.e., using the same approach as KNN).

3.5.4 Random Forests (RF)

The parameters in RF are the splitting variables and split points at each node of the decision tree while the hyper-parameters include the number of trees (B), the number of variables selected randomly at each split (m), and the minimum node size (n_{min}) (Hastie et al., 2009).

In practice, it is generally found that once a sufficient number of trees have been considered in the 'forest', adding additional trees beyond this point does not substantially increase performance (for example, see the large-scale study by Probst & Boulesteix (2018)) and comes at an increased computational cost (Hastie et al., 2009). Further, default values for n_{min} (set to 5 in the ranger R package (Wright & Ziegler, 2017) used to develop the RF models in this study), have also been shown to provide high performance (Díaz-Uriarte & Alvarez de Andrés, 2006). While Probst et al. (2019) confirm that n_{min} is less important to tune than m, they show that at an increased computational cost it can be jointly optimized with m to improve RF performance. However, in order to strike a balance between predictive accuracy and

computational efficiency, n_{min} is fixed in this study (at its default value, 5) and a two stage approach is used to identify suitable values for B and m, as described below.

At first, a preliminary analysis was undertaken revealing that 300 trees ($B = 300$) and a minimum node size of 5 ($n_{min} = 5$) led to stable performance when m was within the range of 1 to 5. Thus, B and n_{min} were held constant at 300 and 5, respectively, and m was optimized through a grid search ($m = 1, 2, \ldots, 5$) using 5-fold cross-validation (where the RMSE was used as the objective function).

The RF models were developed using a combination of the caret and ranger R packages (Kuhn et al., 2019; Wright & Ziegler, 2017).

3.5.5 eXtreme Gradient Boosting (XGB)

The parameters in XGB, similar to RF, are the variables and values used at each split. The XGB hyper-parameters are described in detail in Chen & Guestrin (2016) and Chen et al. (2019) and given below (along with their ranges considered during optimization; integer values are presented with an 'L'):

- nrounds (i.e., the number of trees) (1, 150)
- eta (0.001, 0.5)
- gamma (0, 10)
- max_depth (2L, 12L)
- min_child_weight (1L, 10L)
- subsample (0.5, 1)
- colsample_bytree (0.1, 1)
- lambda (0, 1)
- alpha (0, 1)

The XGB hyper-parameters were optimized using the Bayesian optimization approach of Snoek et al. (2012), which is based on Gaussian Processes. The Bayesian optimization routine adopted the expected improvement method for updating the estimates of the best model parameters (Ribeiro et al., 2020) and was run for 20 iterations after generating 5 initial starting

points for the model parameters. Details on applying Bayesian optimization using the expected improvement method for training DDM is described in Zuo et al. (2020).

The XGB models were built using the ParBayesianOptimization (Wilson, 2019) and xgboost (Chen et al., 2019) R packages.

4 Results

4.1 Uncertainty Intervals: CDDA vs. Benchmark (HM)

Through the CDDA, uncertainty intervals were generated for the ensemble of streamflow simulations using six different DDMs (MLR, KNN, SOV, ANN, RF, XGB). In addition to these six DDMs, one extra variant was considered for both RF and XGB models that consider only the six most important input variables as input to the DDM (by assessing the variable importance scores of RF and XGB). These model variants are referred to as RF_6 and XGB_6. These additional models were created for two reasons: 1) to see if using a smaller number of inputs leads these models to perform poorly (compared to the case when all inputs are considered), and 2) to enable a comparison with the other nonlinear methods (KNN, SOV, and ANN) that, on average, used six input variables as selected by the EA input variable selection method.

Figures 3-5 present the 95% uncertainty intervals of these CDDA variants for three study catchments. These uncertainty intervals result from the parametric uncertainty of the HM model only, i.e. via using multiple (1000) optimized parameter sets for the HBV model, and thus represent only the 95% confidence intervals (95%-CIs). Since a DDM is trained for each set of HM residuals, there are also 1000 DDM, the combination of both HM and DDM, result in the CDDA-based ensemble streamflow simulations. For better visibility, only a short simulation period of 30 days is displayed in Figures 3-5 while a longer simulation period is provided in the Supporting Information. The CDDA-based ensemble streamflow simulations are compared to the benchmark i.e., HM-based ensemble simulated streamflow, which is simply the deterministic output of the HBV model for 1000 optimized model parameter sets. As can be seen from Figures 3-5 (and those in the Supporting Information), the uncertainty intervals of all CDDA and the HBV model are narrow for the three study catchments. The CIs for most of the CDDA variants and HBV at the study catchments do well at covering low flows, moderately well at medium flows, but perform quite poorly at high flows. In addition, the 95 %-CIs for all CDDA closely overlap with the intervals of the benchmark model (HBV). Hence, based only on the visual

assessment, it is difficult to identify which of CDDA variants performs best. To support this analysis, performance metrics were evaluated and are presented in Section 4.2.

The rather narrow uncertainty intervals for all CDDA and the HBV model result from the fact that only parametric uncertainty of the HBV model was considered, whereas other sources of uncertainty (input data, model structure, model output, etc.) were excluded, which may prove useful in improving the quality of the uncertainty intervals. This issue is further discussed in Section 5.2.

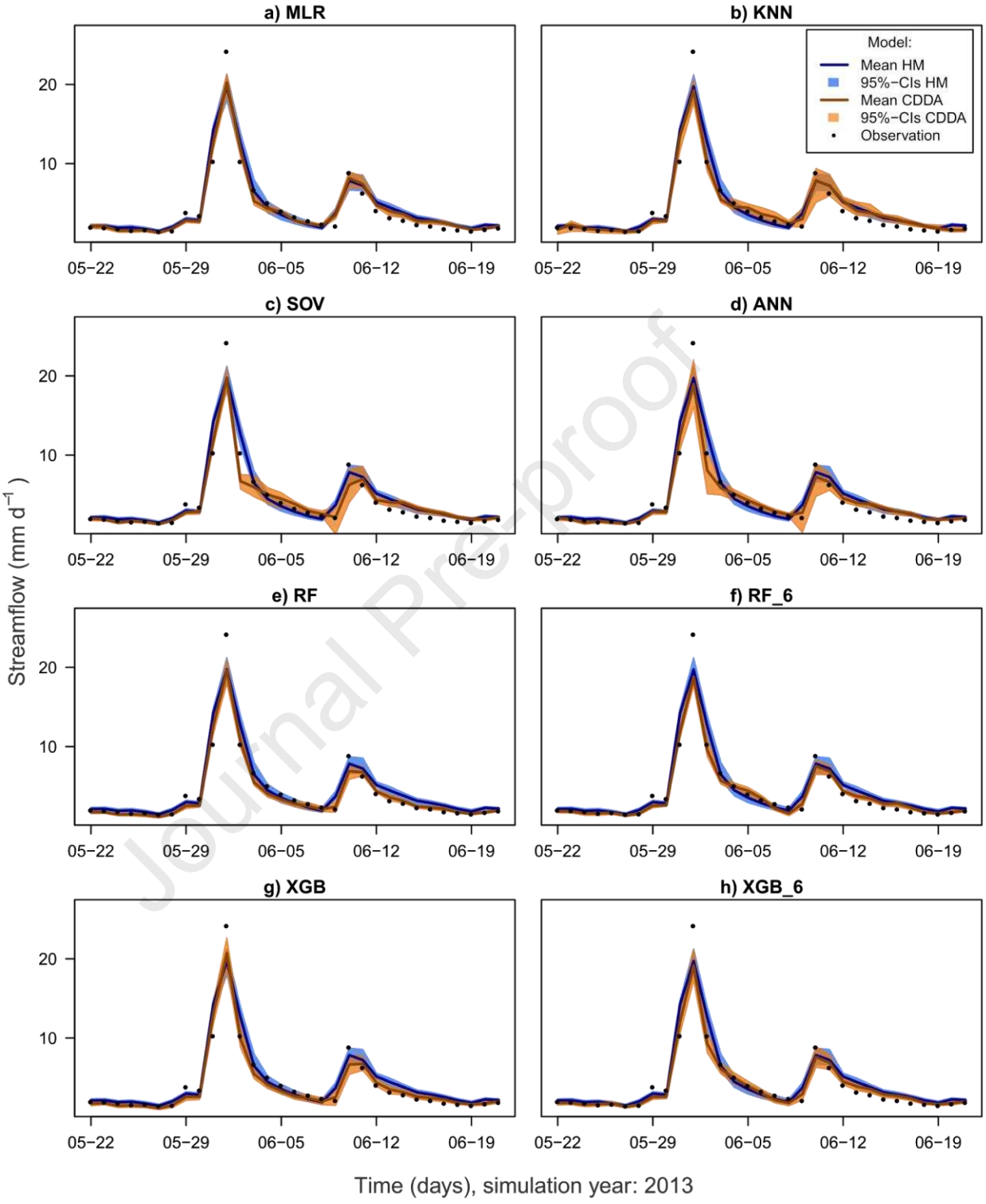


Figure 3. 95% confidence intervals (95%-CIs) of the CDDA using different DDM models versus the benchmark (only HBV model) in the Dünnern catchment (validation period). For longer simulation periods, see Supporting Information.

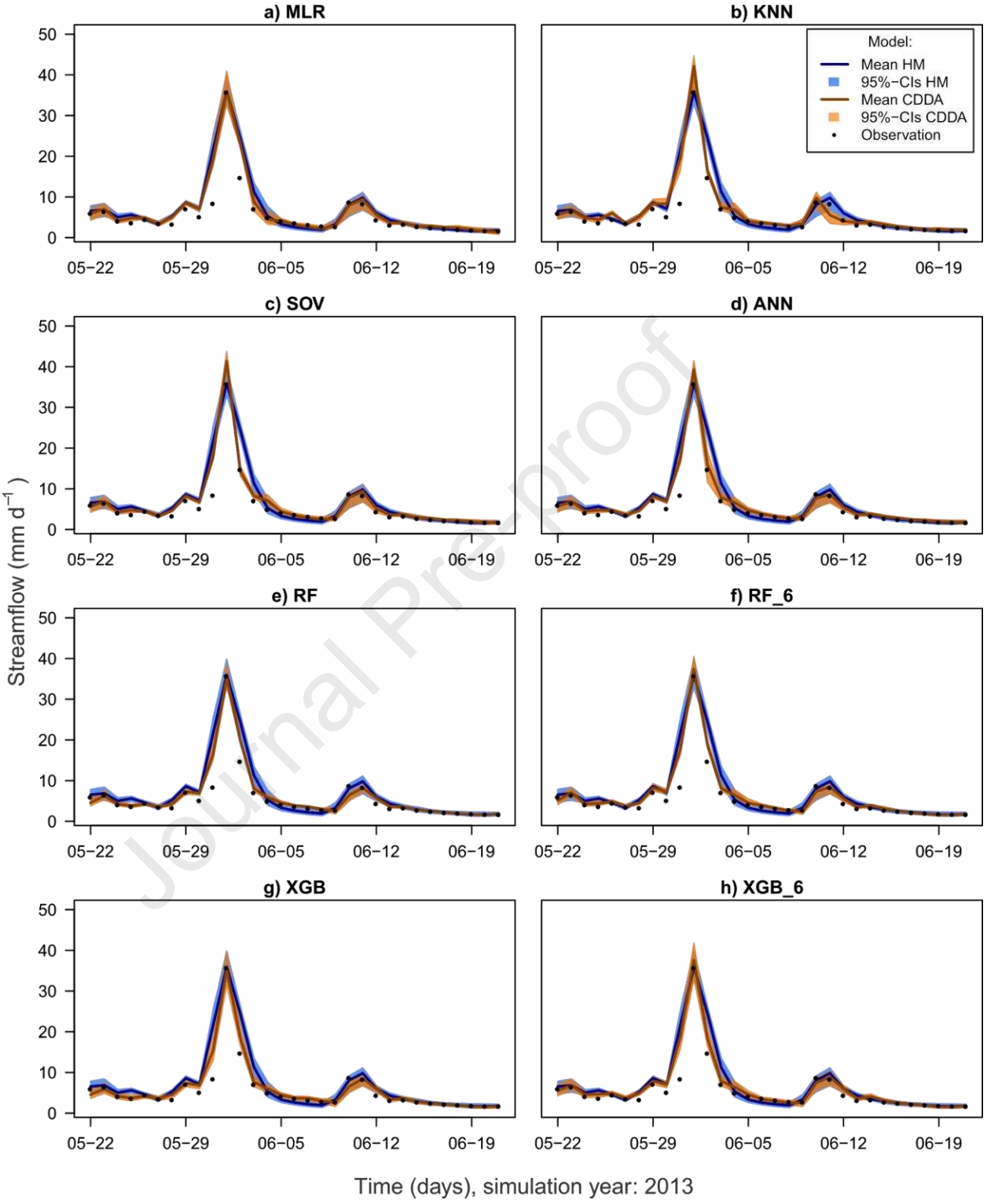


Figure 4. 95% confidence intervals (95%-CIs) of the CDDA using different DDM models versus the benchmark (only HBV model) in the Kleine Emme catchment (during the validation period). For longer simulation periods, see Supporting Information.

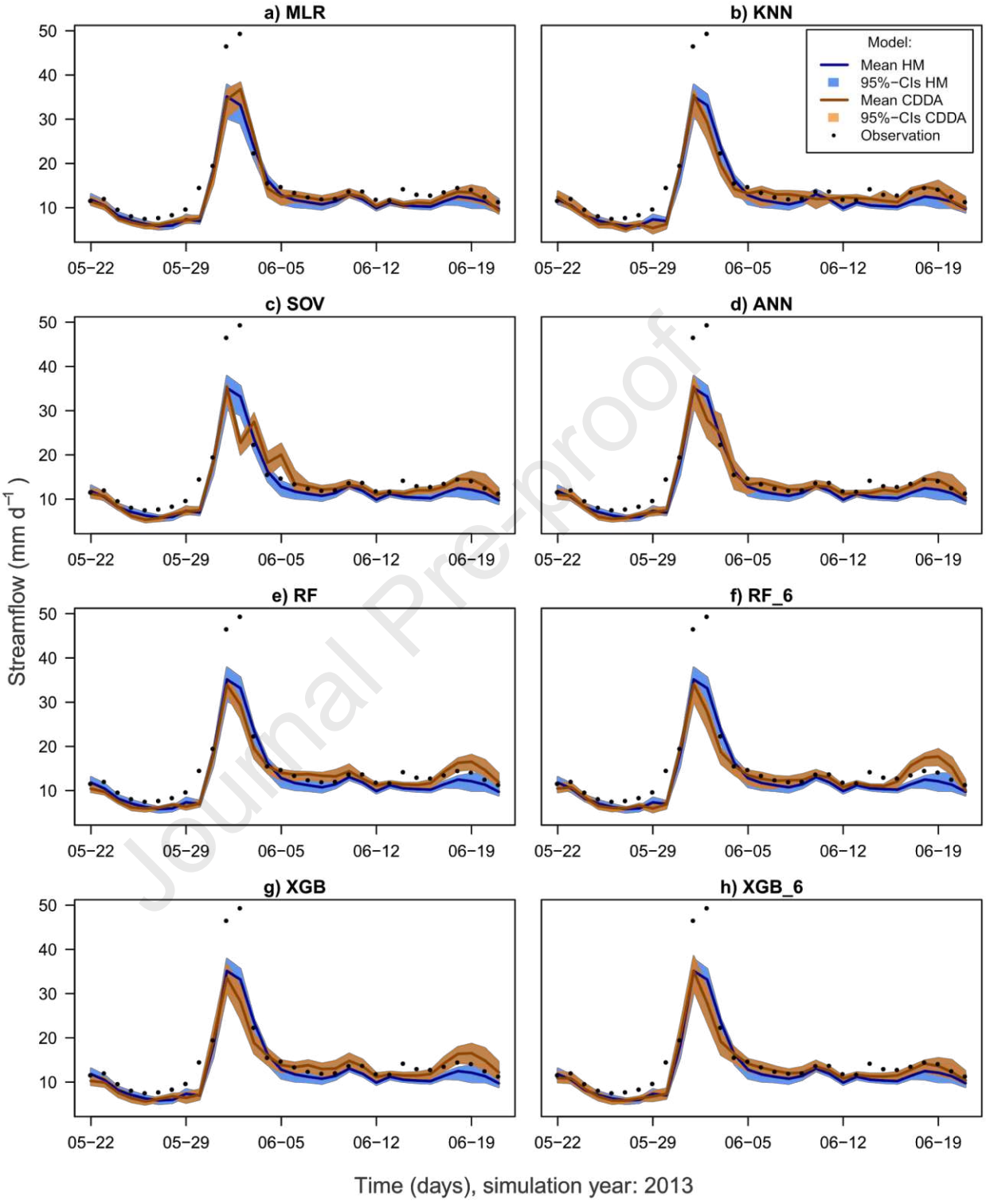


Figure 5. 95% confidence intervals (95%-CIs) of the CDDA using different DDM models versus the benchmark (only HBV model) in the Muota catchment (during the validation period). For longer simulation periods, see Supporting Information.

4.2 Quantitative Assessment of the CDDA Performance

Deterministic performance metrics for the mean ensemble simulation (i.e., the mean over all 1000 ensemble members) are presented for all CDDA variants in Table 2, whereas Table 3 illustrates the ensemble performance metrics (that consider all 1000 ensemble members). These values are compared to the performance criteria computed for the benchmark, which can be used to determine if and by how much the performance of the (ensemble) streamflow simulation is improved when using the CDDA instead of the HM only. The performance criteria are provided for all three study catchments.

Table 2. Deterministic performance metrics for the HM (HBV) and different CDDA variants in the three study catchments. The improvement in the metric value for the CDDA over the benchmark (HBV) is coloured grey.

DDM /Criteria	HM: HBV (benchmark)	MLR	KNN	SOV	ANN	RF	RF_6	XGB	XGB_6
Dünnern									
KGE [-]	0.80	0.79	0.79	0.80	0.80	0.80	0.80	0.80	0.80
NSE [-]	0.76	0.77	0.78	0.79	0.80	0.81	0.80	0.81	0.80
RMSE [mm d^{-1}]	1.07	1.05	1.04	1.01	1.00	0.95	0.98	0.95	0.99
MAE [mm d^{-1}]	0.54	0.51	0.48	0.48	0.45	0.43	0.44	0.42	0.45
Kleine Emme									
KGE [-]	0.87	0.84	0.88	0.85	0.88	0.88	0.88	0.89	0.88
NSE [-]	0.78	0.80	0.81	0.80	0.83	0.85	0.84	0.86	0.84

RMSE [mm d^{-1}]	1.57	1.50	1.46	1.48	1.36	1.28	1.34	1.25	1.34
MAE [m m d^{-1}]	0.83	0.82	0.75	0.75	0.70	0.63	0.66	0.62	0.68
Muota									
KGE [-]	0.87	0.86	0.87	0.87	0.89	0.87	0.89	0.87	0.88
NSE [-]	0.84	0.85	0.84	0.85	0.87	0.87	0.86	0.87	0.86
RMSE [mm d^{-1}]	1.98	1.94	1.96	1.94	1.78	1.82	1.84	1.78	1.83
MAE [mm d^{-1}]	1.10	1.06	1.07	1.03	0.98	0.93	1.00	0.91	0.99

Four deterministic criteria were analyzed: the mean absolute error, root mean square error, Nash Sutcliffe Efficiency, and the Kling-Gupta Efficiency (please refer to the Supporting Information for additional metrics). Optimal performance for NSE and KGE occurs at a value of 1, whereas for RMSE and MAE, optimal performance occurs at a value of 0. Thus, if any CDDA variants lead to an improvement in reference to HBV, NSE, and KGE will increase while RMSE and MAE will decrease. Thus, from Table 2, it can be noticed that most of the CDDA variants lead to an improvement in most metrics over the standalone HBV-based simulations. Namely, NSE, RMSE and MAE were all improved for all DDMs in comparison to the HBV in all three catchments. Regarding KGE, it was improved in the Kleine Emme catchment for models KNN, ANN, RF, RF_6, XGB and XGB_6, and in the Muota catchment for models SOV, ANN, RF, RF_6, XGB and XGB_6. Opposite to that, in the Dünnern catchment, none of the DDMs led to an improvement of the KGE values. Despite the KGE values computed for DDM were slightly smaller, they were still very close to the benchmark value obtained with the HBV model in this catchment. By reviewing the additional performance metrics in the Supporting Information, it

can be seen for the Dünnern catchment that the DDMs have slightly higher bias than HBV, which is the likely reason why the KGE is slightly lower for the DDMs.

Among all tested CDDA variants, RF and XGB led to the largest improvement in the deterministic performance metrics, followed by their variants RF_6 and XGB_6, as compared to the standalone HBV model. For example, RF (RF_6) and XGB (XGB_6) led to improvements in MAE of 20-22% (17-19%), 24-25% (18-20%), and 15-17% (9-10%) for Dünnern, Kleine Emme, and Muota catchments, respectively. While the CDDA based on MLR led to a marginal improvement only in these criteria (e.g., improvements in MAE of 1-6 % across the three catchments). To summarize the effect of the CDDA in terms of its mean ensemble deterministic performance (refer also to the Supporting Information): most DDM are very effective at significantly reducing variance in the resulting simulations (especially, RF, RF_6, XGB, and XGB_6), and in some cases, RF, RF_6, XGB, and/or XGB_6, also improve bias, albeit to a lesser degree (e.g., at Dünnern and Muota catchments).

Given that CDDA generates ensemble streamflow simulations, it is also very important to assess ensemble performance when comparing CDDA to the standalone HBV. To assess the properties of the uncertainty intervals of the CDDA simulations versus the HBV-based simulations (benchmark), three ensemble performance metrics were considered: the mean continuous ranked probability score, the mean interval score, and the average width of uncertainty intervals, which are presented in Table 3.

Table 3. Ensemble performance metrics for the HM (HBV) and different CDDA variants in the three study catchments. The improvement in the metric value for the CDDA over the benchmark (HBV) is coloured grey.

DDM /Criteria	HM: HBV (benchmark)	MLR	KNN	SOV	ANN	RF	RF_6	XGB	XGB_6
Dünnern									
CRPS [mm d^{-1}]	0.48	0.45	0.40	0.41	0.39	0.37	0.39	0.35	0.38

MIS [mm d^{-1}]	13.07	12.19	8.71	10.45	9.52	10.21	10.33	8.07	9.15
AW [mm d^{-1}]	0.55	0.52	0.80	0.58	0.60	0.43	0.48	0.62	0.61
Kleine Emme									
CRPS [mm d^{-1}]	0.70	0.70	0.63	0.64	0.58	0.53	0.56	0.50	0.55
MIS [mm d^{-1}]	16.78	16.92	14.59	15.18	13.42	12.85	13.38	9.88	11.41
AW [mm d^{-1}]	1.19	1.13	1.11	1.07	1.08	0.86	0.95	1.19	1.20
Muota									
CRPS [mm d^{-1}]	0.99	0.95	0.95	0.91	0.85	0.83	0.88	0.77	0.86
MIS [mm d^{-1}]	28.49	27.28	25.77	26.01	23.09	23.59	24.67	17.79	22.00
AW [mm d^{-1}]	1.02	1.02	1.14	0.99	1.08	0.91	1.00	1.38	1.22

Optimal values for CRPS, MIS, and AW should be as small values as possible. Thus, if any of these criteria are lower for CDDA than HBV then it is indicative that CDDA provides superior ensemble performance, with the caveat that lower CRPS and MIS scores are preferred over solely lower AW scores. By evaluating the ensemble performance metrics, it can be noticed that for all three catchments CRPS was decreased for all tested DDMs (with the exception of

MLR for the Dünnern catchment). MIS was decreased for all DDMs in the Dünnern and in the Muota catchments and for most of DDMs in the Kleine Emme catchment (apart from MLR). Regarding AW, this criterion was decreased for most DDMs in Kleine Emme (apart from XGB_6) but only for three models in the other two catchments. The analysis of AW is, however, not straightforward, as the optimal uncertainty intervals should prioritize the smallest values for the other two criteria (i.e. CRPS and MIS) over the AW. Thus, a slightly larger value for AW obtained for a DDM with reference to the benchmark does not necessarily indicate poorer ensemble performance.

Similar to the deterministic performance criteria, when comparing all tested CDDA variants, XGB and RF provided the largest improvement in ensemble performance metrics, followed by their variants RF_6 and XGB_6, with reference to the standalone HBV model. For example, RF (RF_6) and XGB (XGB_6) led to improvements in CRPS of 23-27% (19-21%), 24-29% (20-21%), and 16-22% (11-13%) for Dünnern, Kleine Emme, and Muota catchments, respectively. The CDDA based on MLR led to marginal improvements in these criteria (e.g., CRPS was increased by 6% and 4 % for Dünnern and Muota catchments, respectively, and no change was observed for the Kleine Emme).

In general, based on both deterministic and ensemble performance criteria, it can be concluded that all CDDA variants led to an improvement in the (ensemble) streamflow simulations with reference to the HM-based simulations. Among all tested CDDA variants, the CDDA based on XGB led to the largest improvement in the ensemble streamflow simulations. The second best model was the one based on RF. The third and fourth best models were XGB_6 and RF_6, followed by ANN, SOV and KNN. The worst simulation performance was achieved for the CDDA adopting MLR, which was only slightly better than the benchmark. Thus, it may be argued that linear DDM are inappropriate for fitting HM model residuals in the study catchments (see further Section 5.3).

4.3 Importance of Input Variables in DDMs

To simulate the HM residuals, the different DDMs considered several input variables. These input variables were streamflow, precipitation, and air temperature observed at previous days. In this study, a timeframe of up to 9 days preceding the day of simulation was considered for all three candidate input variables as described in Section 2.4. Hence, the previous 9 days of

streamflow (t-1, ..., t-9), as well as the current and previous 9 days (t-0, ..., t-9) of the precipitation and the air temperature were considered as input variables. In total, 29 different input variables were considered as potential predictors of the HM residuals. IVS was performed for each of the 1000 ensemble members, considering the HM residuals as the target variable, in order to determine the best predictors to use for each individual residual series. Since PCIS and EA are both model-free IVS methods, they were run independent of the DDMs (i.e., the inputs and model parameters were determined separately). In contrast, RF and XGB are model-based IVS approaches; thus, the inputs selected by these methods are related to the model parameters determined during training.

The importance scores of the input variables selected by the different IVS methods associated with the CDDA variants are illustrated in Figures 6-8 for the three study catchments, which summarize the importance scores across all 1000 ensemble members using box plots. The higher the importance score, the stronger impact the input variable has on the simulated residuals. Since the KNN, SOV and ANN models use the exact same inputs as determined by the EA IVS method, only a single plot is considered for these methods. Since PCIS is a linear IVS method, it is solely coupled with MLR. Note that for XGB_6 and RF_6, only six input variables are plotted as these model variants considered only the six most important input variables determined by their base method (XGB and RF). It is important to note that all variables presented in these plots with importance scores above 0 were not necessarily selected for each of the 1000 ensemble members, but any inputs with importance scores higher than 0, were selected at least once (in the 1000 ensemble members).

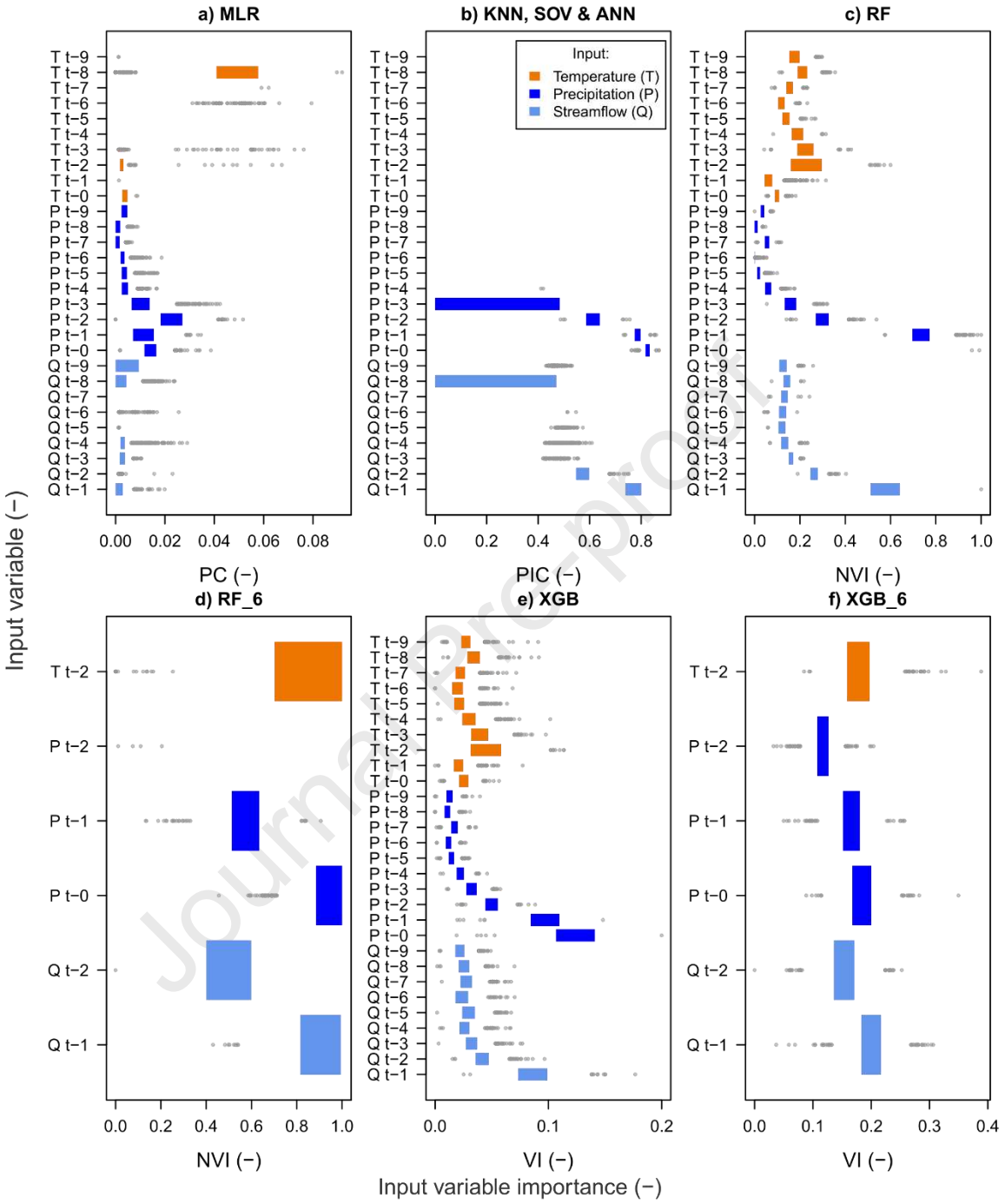


Figure 6. Importance of input variables in different DDMs (Dünnern). Different variable importance scores for the DDM are described in Section 2.4. ***Q***, ***P*** and ***T*** represent observed streamflow, precipitation and air temperature at preceding days (t-0, t-1, ..., t-9). Note that for observed streamflow at t-0 is not considered as it is the target simulation day.

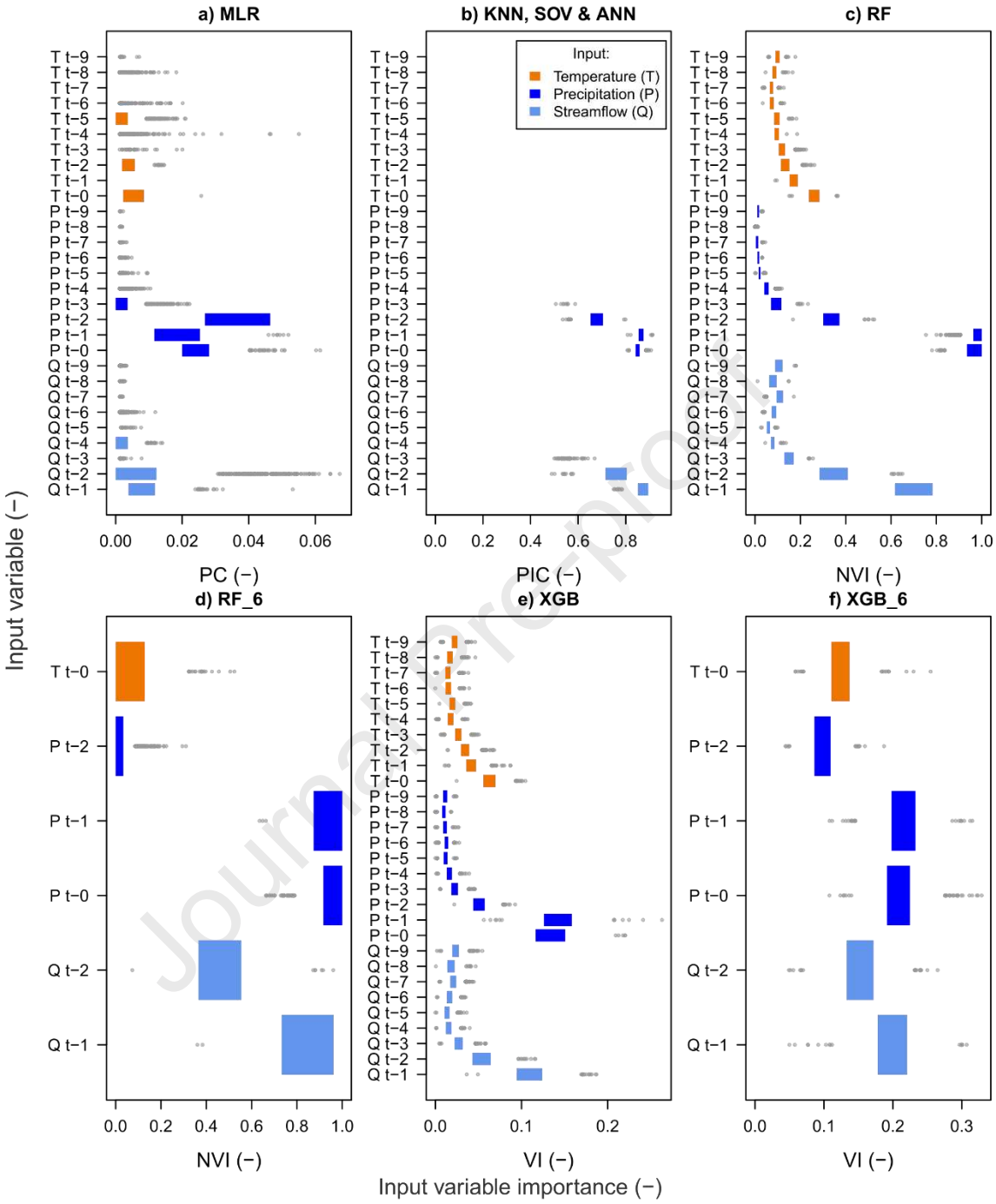


Figure 7. Importance of input variables in different DDMs (Kleine Emme). See Figure 6 for additional information.

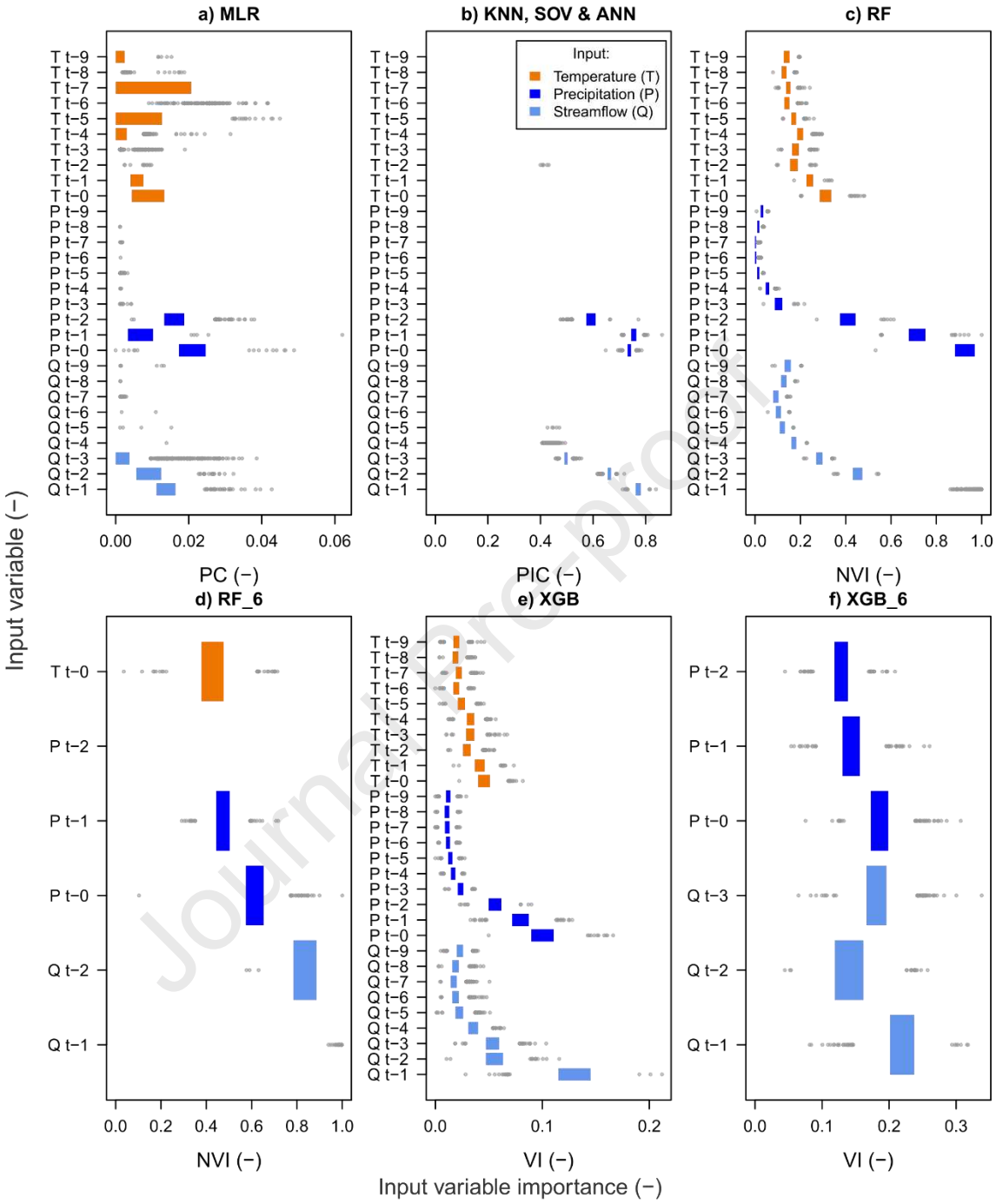


Figure 8. Importance of input variables in different DDMs (Muota). See Figure 6 for additional information.

From Figures 6-8 it can be seen that, generally, the importance of the input variables increases with decreasing the lag time, i.e., a higher importance is given to input variables whose

lag time directly precedes the simulation day. For all CDDA variants, among the three considered input variables, precipitation was the strongest predictor of the HM residuals, followed by streamflow. The air temperature was the weakest predictor among all DDMs apart from the MLR model where it was a stronger predictor than the streamflow in two out of three catchments.

Looking at different models in detail, it appears that for KNN, SOV and ANN (which use the EA method based on conditional mutual information for IVS), the air temperature does not play any significant role while the importance of the streamflow from the three preceding days (t-1, ..., t-3) and the precipitation from the three preceding days, including the day of simulation, (t-0, ..., t-3) were the most important for simulating HM residuals in all three catchments. For MLR, the precipitation (t-0, ..., t-3) was very important in all catchments, whereas the importance of the air temperature and streamflow varied depending on the catchment. In RF and XGB, the precipitation (t-0, ..., t-3), the streamflow (t-1, ..., t-3) and the temperature (t-0, ..., t-3) were all very important in all three catchments with precipitation (t-0, t-1, t-2) and streamflow t-1 being the most important. Yet, it can be noticed that all input variables have above 0 importance in RF and XGB, meaning that all variables contribute to the overall ensemble HM residual simulations. For RF_6 and XGB_6 that consider only the six most important variables (from their respective base model), the selected variables were always streamflow t-1 and t-2, and precipitation t-0, t-1, and t-2 in all three catchments. The sixth most important selected variable varied depending on the catchment and it was either air temperature t-2 (Dünnern), air temperature t-0 (Kleine Emme), or streamflow t-3 (Muota). The order of importance for these six variables varied depending on the catchment; however, streamflow t-1 and precipitation t-0 and t-1 were always found to be very important. An interesting result was obtained by RF_6 for the Muota catchment: when only the six most important inputs identified by RF were considered in RF_6, it was found that precipitation (t-2) did not add any information that was useful when simulating the HM residuals. This suggests that, when all 29 inputs are considered, there is another input (outside of the other five selected inputs), that, when combined with precipitation (t-2), adds information that is useful for simulating the HM-residuals. While it is outside the scope of this research to identify the other inter-dependent input, the interested reader may refer to Galelli et al. (2014) whom discuss the inter-dependency of inputs in IVS.

4.4 Effect of the Ensemble Size

Simulations of both CDDA and HM models were based upon 1000 ensemble members that originate from the 1000 optimized parameter sets of the HBV model. Yet, it is not clear whether use of all 1000 ensemble members is of a value for the CDDA. As it appears from Figures 3-5, the uncertainty intervals are rather narrow, which implies that some members may be redundant. This issue was investigated here by exploring the effect of the ensemble size on the simulation performance. For this purpose, Figure 9 can be used to analyze how the CRPS changes as the ensemble member size grows for all CDDA variants and the benchmark model (HBV) in the three study catchments.

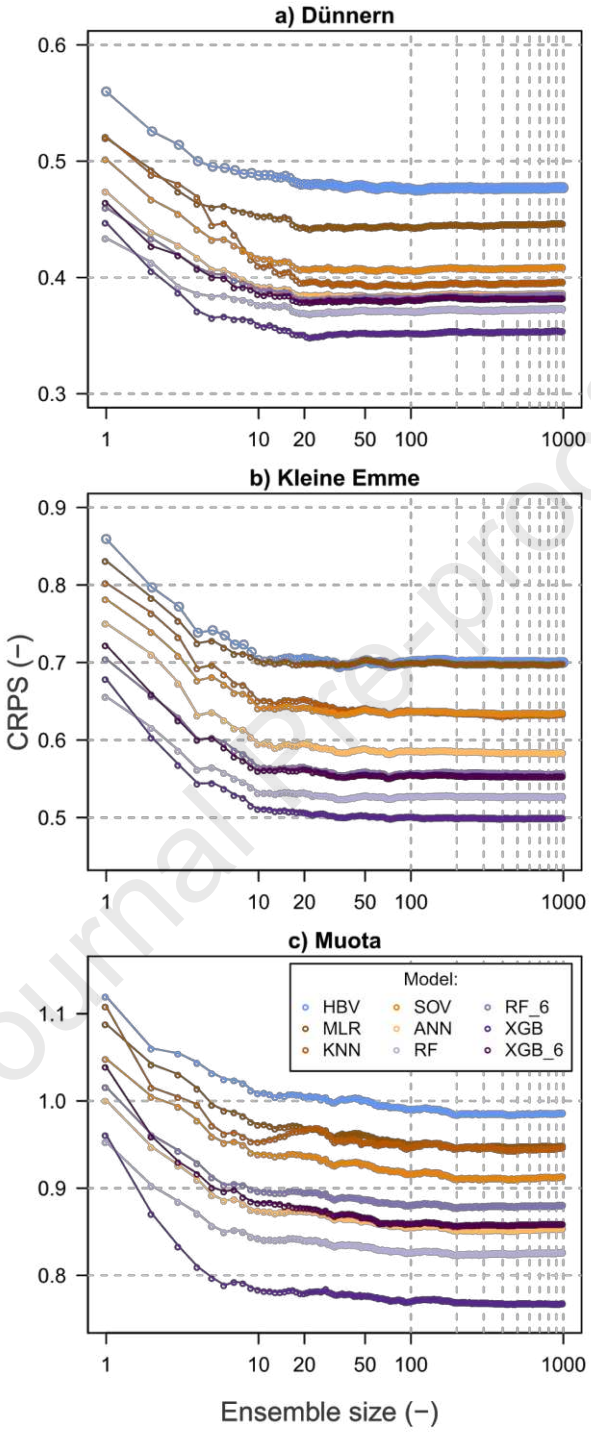


Figure 9. Mean continuous ranked probability score (CRPS) as a function of the ensemble size for different DDMs and the benchmark (HBV) in the three catchments (note: the log scale for the ensemble member size).

Based on Figure 9, it can be seen that as few as 100 ensemble members provide roughly the same ensemble performance as 1000 ensemble members for both the HBV-based simulations (benchmark) and for the CDDA-based simulations when considering the CRPS. Using less than 100 ensemble members leads to a visible drop in the CRPS, whereas using more than 100 ensemble members does not significantly improve the model performance. The strongest improvement is noticed between one and 10 ensemble members, which seems logical when moving from the deterministic approach (only one ensemble member) towards a probabilistic approach (several ensemble members). This effect of growing the ensemble size on the model performance was visible for all tested CDDA variants as well as for the HBV model in all three study catchments.

5 Discussion

In this work an ensemble-based conceptual-data-driven approach (CDDA) was developed that consists of a hydrological model (HM) to simulate precipitation-streamflow processes and a data driven model (DDM) to model residuals of the HM. The goal of CDDA is to improve the predictive capability of the ensemble streamflow simulation compared to the HM. Results from three study catchments in Switzerland have demonstrated that coupling DDMs with a simple hydrological model (HBV) is capable of improving the predictive performance of the HM, assessed by both the deterministic and ensemble performance criteria, as compared to the standalone HM (benchmark). Below the major findings of this work are explored in detail, its limitations are discussed, and some recommendations for future research are given.

5.1 Different DDM and Importance of Input Variables

Six different DDMs and two additional variants were tested for simulating HM residuals. These DDMs link the HM residuals on the simulation day (t-0) to observed precipitation and air temperature on the simulation day and those preceding it (t-0 to t-9) as well as streamflow on preceding days (t-1 to t-9). This maximal lag time of nine days was determined by studying the nonlinear correlation between explanatory variables and the HM residuals (see Section 3.5). These results demonstrated that, generally, all DDMs coupled with the HM within the CDDA framework led to an improvement in the ensemble streamflow simulations with reference to the HM-based simulations (Section 4.2). This increase in the streamflow simulation performance (measured by deterministic and ensemble performance metrics) was, however, not the same for

all models and for the simplest DDM (i.e. MLR) it was only marginal. The largest increase in the model performance was achieved with XGB and then with RF models. The XGB and RF consistently had the best performance across all three catchments.

Analysis of the input variable importance (Section 4.3) revealed that the importance of the input variables increases with decreasing lag time, i.e., input variables from days directly preceding the simulation day were of a higher importance than those from the distant past. For all DDMs, the observed streamflow (t-1, …, t-3) and precipitation (t-0, …, t-3) from the current and preceding three days had the highest impact on the HM simulated residuals while for certain methods (RF and XGB), the air temperature from the current and three preceding days (t-0, …, t-3) were also deemed useful inputs. However, for XGB and RF, observations farther in the past (t-4 to 4-9) also had an impact on the simulated residuals. Hence, it appears that, despite the maximal lag time of nine days, the effective memory length, which determines the lag length beyond which input variables have only marginal or null effect on the simulated output, appears to be about three days directly preceding the simulation day. This memory lag length (Bowden et al., 2005) may be further explained by the catchment memory to past inputs that determines how long water is retained in the catchment in different forms such as aquifers, snowpack or groundwater storage (Müftüoğlu, 1984; Rajurkar et al., 2004). This catchment memory is often identified as precipitation influence history (Müftüoğlu, 1991), which is the strongest predictor of the streamflow. The length of the catchment memory varies between catchments and may be from several hours to several days and longer. As the results indicate, not only the simulated streamflow but also residuals of the hydrological model may be linked to the catchment memory. From these findings, it appears that the more complex DDMs, that link the residuals to additional input variables (i.e. use longer lag times), have better predictive skill in correctly mimicking the residuals of the HM. Hence, although the most important input variables seem to be observations from the last three days, using longer lag times than that with more complex models further improves the model performance. Note, however, that using longer lag times with simpler models does not improve the model performance.

Next, among the three considered input variables, the observed precipitation was the strongest predictor of the HM residuals, followed by the observed streamflow, while the observed air temperature was the weakest predictor of residuals. This seems reasonable as the precipitation is usually also the strongest predictor of streamflow in any HM that relies on the

precipitation-streamflow generation concept. The air temperature is usually used within many HM to determine the form of the precipitation (liquid or solid) and to determine whether snow melt occurs (Seibert & Vis, 2012). Thus, it seems reasonable that its impact is smaller than that of the precipitation, which determines the amount of water entering the catchment. The importance of the observed streamflow as a predictor of model residuals can be explained by the auto-correlation effect present for most hydrological models, i.e., when streamflow residuals at t-0 are correlated with residuals at preceding time steps (t-1, t-2, …, etc.) (Sorooshian & Dracup, 1980; Yang et al., 2007; Sikorska et al., 2012). The strength of this auto-correlation likely depends on the precipitation and flow conditions and for wet or high flow periods it is expected to be higher than for dry or low flow periods. This concept has been explored by Del Giudice et al. (2013), who linked the residuals of a hydrological model to the precipitation amount or streamflow via Bayesian inference. The authors found that, if conditioned on one variable, the streamflow-dependent description of model residuals was better-performing than the precipitation-dependent error. However, this seems opposite to findings from this study; however, DDMs that rely on only a single input variable were not investigated here, instead DDMs always included several different input variables.

5.2 Applicability of the Ensemble-based CDDA Approach and Limitations

The results demonstrated that the ensemble-based CDDA framework is very promising for simulating HM residuals and generating an improved ensemble of streamflow simulations when compared to the HM benchmark. Although, in this study, only a single conceptual model (HBV) was applied to simulate the precipitation-streamflow process. However, the CDDA is not limited to this model and any other hydrological model can be coupled with the proposed framework. It should be noted that using any other HM or even an HBV model with different parameters, would require the DDMs to be recalibrated since the HM will generate different residuals. In addition, since different HMs may use other input variables (e.g., potential evapotranspiration), the importance of those new variables should be also explored to develop the best possible DDM for the given dataset. The same holds for an application of the proposed approach to other sites, where the HM residuals may exhibit different properties than those obtained in this study. Hence, the DDMs should be re-calibrated (trained) based on local observations. Yet, the CDDA, as a general framework for improving ensemble-based hydrological model simulations, remains essentially the same for different hydrological models

or study sites. In general, the CDDA can be used to identify the best DDM and select the most suitable input variables for a given location.

Using a DDM for simulating HM residuals is highly advantageous since it does not require any explicit assumptions on the characteristics of the residuals as is the case in all Bayesian-based methods. Thus, there is no need to assume the residuals' independence, normality, or auto-correlation structure. Moreover, with methods such as RF and XGB, it appears there is little need to perform pre-selection of input variables, as important variables tend to be included in the modelling framework (as corroborated by model-free IVS methods, such as EA), allowing the model to distinguish (on its own) relevant from irrelevant and/or redundant inputs. However, the use of inputs that have little justifiable use in the model is not advocated, as it is rational to only include in the DDMs input variables that have an impact on the HM model residuals, and thus streamflow. Even for methods such as RF and XGB, performing IVS may lead to improved performance and models with lower complexity (Tyralis & Papacharalampous, 2017; Hadi et al., 2019). Indeed, for methods such as MLR, KNN, SOV, and ANN, input variable selection is a necessity, since the models on their own lack the ability to 'filter out' non-useful inputs (Galelli et al., 2014). Although, recent research has made progress in addressing this short-coming by placing tunable weights on the input variables used in the DDM , which may be tuned simultaneously along with the DDM parameters to provide insights on how the input variables impact model predictions (Yang et al., 2020b). Further, the use of IVS can also help reduce the computational burden of model development, which is substantial for large ensembles (e.g., 1000 members), although running the models is extremely quick in an operational setting. Thus, by reducing the computational burden, IVS may also increase the exploration of alternative DDMs.

Additionally, IVS may also prove useful in reducing the number of ensemble members in the CDDA by identifying a reduced number of members that maintain a similar level of performance as the initial ensemble size. For example, it was found that 100 ensemble members (i.e., the first 100 out of 1000 randomly generated members) carried roughly the same level of information as the entire 1000-member ensemble (see Section 4.4).

The major limitation of this study is that only parametric uncertainty of the hydrological model (here, HBV) was investigated and represented with multiple optimized parameter sets

(1000) without explicitly considering any other uncertainty sources of the HM, such as uncertainties in the inputs or model structure (Renard et al., 2011; Sikorska & Renard, 2017). Hence, only confidence intervals of streamflow simulations can be computed while prediction intervals cannot be provided. This can also explain why the uncertainty intervals are very narrow for all simulations, i.e. for the HM alone as well as for different CDDA variants. Note that residuals simulated with a DDM represent the remaining uncertainty of the hydrological model that is not explicitly considered. Yet, as the output of the CDDA is conditioned on the simulations from the HM, narrow uncertainty intervals resulting from the HM translate to narrow uncertainties resulting from the CDDA. However, in order to focus on the exploring the difference in performance across several DDMs, different uncertainty sources (i.e., other than parameter uncertainty) were explicitly not considered here. The effect of other uncertainty sources should be, however, investigated in future studies.

At present, the proposed ensemble-based CDDA framework has only been tested in three gauged catchments. Yet, it would be very interesting to test the approach in ungauged catchments (without streamflow observations). This would enable for improving streamflow simulations at sites where performance is the lowest (due to a lack of recorded data to calibrate a HM). As a direct calibration of the CDDA and its components (i.e., the HM or DDM) is not possible at ungauged locations, other methods to inform both models should be searched for: regionalization approaches could be used to inform the HM model parameters; while training in a large set of catchments of different properties could help to constrain information on simulated residuals, and transfer these residuals to the ungauged catchments that have similar properties. Gauch et al. (2021) have revealed that using data from a large set of catchments to train a single DDM yields better streamflow simulation results, also in poorly gauged regions, suggesting potential towards generalization of DDMs. Wu et al. (2019) have also demonstrated that a DDM used to simulate residuals of an un-calibrated HM may improve its predictive performance. Finally, it is important to note that in an operational context, if there are no streamflow observations available, then lagged measurements of the streamflow cannot be used as an input variable for simulating streamflow via the CDDA. This is relevant, as there may be cases where streamflow gauging stations may be temporarily offline or are no longer in operation due to decommissioning, damage due to a flooding event, etc. (Tencaliec et al., 2015; Villalba et al., 2021); thus, obtaining updated information on recent streamflow is not possible. In these cases,

while streamflow data may be used for calibrating the CDDA (i.e., using streamflow measurements from a limited historical database as the target variable) other input variables may be required instead of streamflow to improve its predictive performance.

5.3 Recommendations

To summarize, it appears from the results that it is generally better to include (almost) any (nonlinear) DDM to simulate the residuals of a hydrological model than using a standalone HM. However, in the case of the simplest data-driven model (MLR), the improvement in ensemble streamflow simulation was only marginal, whereas more complex (nonlinear) models (XGB and RF) led to a significant improvement in the ensemble streamflow simulation. Moreover, use of a DDM to simulate residuals is linked with some additional computational efforts. Thus, the selection of DDM is very important when computational power is restricted. Since it was found that MLR barely improved the original HM simulations, this approach is not recommended, unless, of course, there is reason to believe the relationship between HM residuals and input variables is linear, which did not appear to be the case for the study catchments. However, it was found that XGB and RF both led to significant gains in deterministic and ensemble performance over the standalone HM; given that both methods inherently perform IVS, and thus, require little user intervention while providing impressive performance, are recommended for further study. Additionally, since it was found that only 100 ensemble members provided a similar level of performance as the initial 1000-member ensemble, it is plausible, although it was not verified, that a smaller and carefully selected set of ensemble members, identified via IVS, may provide a similar level of performance as the initial ensemble size. This could be explored in two ways, by performing IVS: 1) on the HM ensemble and then developing DDMs for this reduced set or 2) on the CDDA ensemble. Speculation as to which of the two approaches would result in better performance is out of the scope of this paper but it is recommended as an interesting line of future research. Additionally, it is recommended to include other sources of uncertainty in the ensemble-based CDDA, such as those related to inputs, input variable selection, parameters, and model output, in order to improve uncertainty estimation and the overall utility of the probabilistic forecasts. Finally, other potentially useful input variables, such as potential evapotranspiration, soil moisture, relative humidity, wind speed, are recommended to be explored in the DDM, even if such methods are unable to be included in a particular HM.

6 Conclusions

A novel ensemble-based conceptual-data-driven approach (CDDA) has been proposed. It consists of a hydrological model (HM) to simulate an ensemble of precipitation-streamflow processes and a data-driven model (DDM) to model an ensemble of HM residuals. The DDM takes precipitation, air temperature, and streamflow observed at preceding days to simulate the residuals. Such a CDDA combines the advantages of a HM, respecting hydrological processes, with the ability of the DDM to simulate complex (nonlinear) relationships between input-target (explanatory-response) variables by tackling auto-correlated HM residuals. The CDDA does not require any statistical assumptions on the model residuals and it provides a framework for identifying suitable DDMs and input variables. Moreover, the ensemble-based CDDA is very flexible as it can be coupled with any hydrological model and any DDM for ensemble streamflow simulation. The selection of potential input variables can also be adjusted based on available data as well as user needs and specific conditions (e.g., hydrological model or type of runoff generation). Generally, the CDDA has been shown to be a very promising approach to improve ensemble streamflow simulations. Among eight variants of different DDMs, eXtreme Gradient Boosting (XGB) and Random Forests (RF) were found to be the most accurate predictors of the HM residuals and also required less user intervention in terms of selecting appropriate inputs to the DDM. Also, precipitation and streamflow from the three days directly preceding the simulation day were found to have the largest impact on simulated residuals. Based on the results, XGB and RF models are recommended to simulate the residuals of the hydrological model within the ensemble-based CDDA framework. It was also found in this study that the number of ensemble members may be substantially reduced, i.e., from 1000 to 100, without significantly affecting model performance; this is especially important for cases where computational resources are limited.

Acknowledgments

The authors wish to thank the three anonymous reviewers who provided constructive feedback that has helped to improve the quality of this paper. This research did not receive funding from any agency and was carried out according to the authors' curiosity.

The Swiss Federal Office for the Environment (FOEN) is acknowledged for providing the streamflow data used in this study that were available from the project No. 15.0054.PJ/O503-

1381. Calibration of the HBV model was run using the ScienceCloud provided by S3IT at the University of Zurich.

Data availability

The observed discharge data for calibrating the hydrologic model can be ordered from the FOEN (https://www.bafu.admin.ch, last access: 11 December 2020), the observed meteorological data from MeteoSwiss (http://www.meteoswiss.ch, last access: 11 December 2020). The latest version of the HBV model is available at: https://www.geo.uzh.ch/en/units/h2k/Services/HBV-Model.html

Funding:

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

Abba, S. I., Jasim Hadi, S., Sammen, S. S., Salih, S. Q., Abdulkadir, R. A., Bao Pham, Q., & Mundher Yaseen, Z. (2020). Evolutionary computational intelligence algorithm coupled with self-tuning predictive model for water quality index determination. *Journal of Hydrology*, 587, 124974. https://doi.org/10.1016/j.jhydrol.2020.124974

Abrahart, R. J., Anctil, F., Coulibaly, P., Dawson, C. W., Mount, N. J., See, L. M., et al. (2012). Two decades of anarchy? Emerging themes and outstanding challenges for neural network river forecasting. *Progress in Physical Geography*, 36(4), 480–513, https://doi.org/10.1177/0309133312444943.

Altman, N. S. (1992). An introduction to kernel and nearest-neighbor nonparametric regression. The American Statistician. 46 (3): 175–185. doi:10.1080/00031305.1992.10475879.

Althoff, D., Bazame, H. C., Garcia, J. N. (2021) Untangling hybrid hydrological models with explainable artificial intelligence. H2Open Journal 2021; h2oj2021066. doi:https://doi.org/10.2166/h2oj.2021.066.

Amorocho, J. (1963). Measures of the linearity of hydrologic systems. *Journal of Geophysical Research*, 68(8), 2237–2249. https://doi.org/10.1029/JZ068i008p02237

Amorocho, J., & Brandstetter, A. (1971). Determination of Nonlinear Functional Response Functions in Rainfall-Runoff Processes. *Water Resources Research*, 7(5), 1087–1101. https://doi.org/10.1029/WR007i005p01087

Belgiu, M., & Drăgu, L. (2016). Random forest in remote sensing: A review of applications and future directions. *ISPRS Journal of Photogrammetry and Remote Sensing*, 114, 24–31. https://doi.org/10.1016/j.isprsjprs.2016.01.011

Beven, K. & Freer, J. (2001) Equifinality, data assimilation, and uncertainty estimation in mechanistic modelling of complex environmental systems using the GLUE methodology, *Journal of Hydrology*, 249, 11-29, https://doi.org/10.1016/S0022-1694(01)00421-8.

Boucher, M.-A., Quilty, J., & Adamowski, J. (2020). Data assimilation for streamflow forecasting using Extreme Learning Machines and Multilayer Perceptrons. *Water Resources Research*, 56(6), e2019WR026226. https://doi.org/10.1029/2019WR026226.

Bergström, S., & Forsman, A. (1973). Development of a conceptual deterministic rainfall-runoff model. *Nordic Hydrology*, 4(3), 147–170, https://doi.org/10.2166/nh.1973.0012.

Beygelzimer, A., Kakade, S., & Langford, J. (2006). Cover Trees for Nearest Neighbor. In *Proceedings of the 23rd International Conference on Machine Learning,* pp. 97–104. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/1143844.1143857Beygelzimer, A., Kakadet, S., Langford, J., Arya, S., Mount, D., & Li, S. (2019). FNN: Fast Nearest Neighbor Search Algorithms and Applications. Retrieved from https://cran.r-project.org/package=FNN

Bhagat, S. K., Tiyasha, T., Awadh, S. M., Tung, T. M., Jawad, A. H., & Yaseen, Z. M. (2021). Prediction of sediment heavy metal at the Australian Bays using newly developed hybrid artificial intelligence models. *Environmental Pollution*, 268, 115663. https://doi.org/10.1016/j.envpol.2020.115663

Biau, G. (2012). Analysis of a random forests model. *The Journal of Machine Learning Research*, 13(1), 1063–1095.

Bowden, G. J., Dandy, G. C., & Maier, H. R. (2005) Input determination for neural network models in water resources applications. Part 1 - background and methodology, *Journal of Hydrology*, 301, 1-4, 75-92, https://doi.org/10.1016/j.jhydrol.2004.06.021.

Boucher, M. A., Anctil, F., Perreault, L., & Tremblay, D. (2011). A comparison between ensemble and deterministic hydrological forecasts in an operational context. *Advances in Geosciences*, 29, 85–94. https://doi.org/10.5194/adgeo-29-85-2011

Breinl, K., (2016) Driving a lumped hydrological model with precipitation output from weather generators of different complexity. *Hydrological Sciences Journal*, 61, 1395–1414. doi:10.1080/02626667.2015.1036755

Brédy, J., Gallichand, J., Celicourt, P., & Gumiere, S. J. (2020). Water table depth forecasting in cranberry fields using two decision-tree-modeling approaches. *Agricultural Water Management*, 233, 106090. https://doi.org/10.1016/j.agwat.2020.106090

Breiman, L. (1996). Bagging predictors. *Machine Learning*, 24(2), 123–140. https://doi.org/10.1007/BF00058655

Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324

Bröcker, J. (2012). Evaluating raw ensembles with the continuous ranked probability score. *Quarterly Journal of the Royal Meteorological Society*, 138(667), 1611–1617. https://doi.org/10.1002/qj.1891

Brown, G., Pocock, A., Zhao, M.-J., & Lujan, M. (2012). Conditional Likelihood Maximisation: A Unifying Framework for Mutual Information Feature Selection. *Journal of Machine Learning Research*, 13, 27–66. https://doi.org/10.1016/j.patcog.2015.11.007

Chen, C., Zhang, Q., Yu, B., Yu, Z., Lawrence, P. J., Ma, Q., & Zhang, Y. (2020). Improving protein-protein interactions prediction accuracy using XGBoost feature selection and stacked ensemble classifier. *Computers in Biology and Medicine*, 123, 103899. https://doi.org/10.1016/j.compbiomed.2020.103899

Chen, H., Liu, X., Jia, Z., Liu, Z., Shi, K., & Cai, K. (2018). A combination strategy of random forest and back propagation network for variable selection in spectral calibration. *Chemometrics and Intelligent Laboratory Systems*, 182, 101–108. https://doi.org/10.1016/j.chemolab.2018.09.002

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794. New York, NY, USA: ACM. https://doi.org/10.1145/2939672.2939785

Chen, T. T. He, M. Benesty, V. Khotilovich, Y. Tang, H. Cho, K. Chen, R. Mitchell, I. Cano, T. Zhou, M. Li, J. Xie, M. Lin, Y. Geng, & Y. Li. xgboost: Extreme Gradient Boosting, 2019. URL https://CRAN.Rproject.org/package=xgboost. R package version 0.90.0.2.

Gneiting, T., & Raftery, A. E. (2007). Strictly Proper Scoring Rules, Prediction, and Estimation. *Journal of the American Statistical Association*, 102(477), 359–378. https://doi.org/10.1198/016214506000001437

Chen, X., & Ishwaran, H. (2012). Random forests for genomic data analysis. *Genomics*, 99(6), 323–329. https://doi.org/10.1016/j.ygeno.2012.04.003

Cheng, B., & Titterington, D. M. (1994). Neural Networks: A Review from a Statistical Perspective. *Statistical Science*, 9(1), 2–30. Retrieved from http://www.jstor.org/stable/2246275

Chivers, B. D., Wallbank, J., Cole, S. J., Sebek, O., Stanley, S., Fry, M., & Leontidis, G. (2020). Imputation of missing sub-hourly precipitation data in a large sensor network: A machine learning approach. *Journal of Hydrology*, 588, 125126. https://doi.org/10.1016/j.jhydrol.2020.125126

Del Giudice, D., Honti, M., Scheidegger, A., Albert, C., Reichert, P., & Rieckermann, J. (2013) Improving uncertainty estimation in urban hydrological modeling by statistically describing bias, *Hydrology and Earth System Sciences*, 17, 4209–4225, https://doi.org/10.5194/hess-17-4209-2013.

Deng, H., & Runger, G. (2013). Gene selection with guided regularized random forest. *Pattern Recognition*, 46(12), 3483–3489. https://doi.org/10.1016/j.patcog.2013.05.018

Díaz-Uriarte, R., & Alvarez de Andrés, S. (2006). Gene selection and classification of microarray data using random forest. *BMC Bioinformatics*, 7, 1–13. https://doi.org/10.1186/1471-2105-7-3

Diskin, M. H., Boneh, A., & Golan, A. (1984). Identification of a Volterra series conceptual model based on a cascade of nonlinear reservoirs. *Journal of Hydrology*, 68(1), 231–245. https://doi.org/10.1016/0022-1694(84)90213-0

Ebrahimi, E., & Shourian, M. (2020). River Flow Prediction Using Dynamic Method for Selecting and Prioritizing K-Nearest Neighbors Based on Data Features. *Journal of Hydrologic Engineering*, 25(5), 04020010. https://doi.org/10.1061/(asce)he.1943-5584.0001905

Ehlers, L. B., Wani, O., Koch, J., Sonnenborg, T. O., & Refsgaard, J. C. (2019). Using a simple post-processor to predict residual uncertainty for multiple hydrological model outputs. *Advances in Water Resources*, 129, 16–30. https://doi.org/10.1016/j.advwatres.2019.05.003

Fahimi, F., Yaseen, Z. M., & El-shafie, A. (2017). Application of soft computing based hybrid models in hydrological variables modeling: a comprehensive review. *Theoretical and Applied Climatology*, 128(3), 875–903. https://doi.org/10.1007/s00704-016-1735-8

Fan, J., Yue, W., Wu, L., Zhang, F., Cai, H., Wang, X., et al. (2018). Evaluation of SVM, ELM and four tree-based ensemble models for predicting daily reference evapotranspiration using limited meteorological data in different climates of China. *Agricultural and Forest Meteorology*, 263, 225–241. https://doi.org/10.1016/j.agrformet.2018.08.019

Farmer, W. H., & R. M. Vogel (2016), On the deterministic and stochastic use of hydrologic models, *Water Resources Research*, 52, 5619–5633, doi:10.1002/2016WR019129.

Fatehi, I., Amiri, B. J., Alizadeh, A., & Adamowski, J. (2015). Modeling the relationship between catchment attributes and in-stream water quality. *Water Resources Management*, 29(14), 5055–5072. https://doi.org/10.1007/s11269-015-1103-y

Fawagreh, K., Gaber, M. M., & Elyan, E. (2014). Random forests: From early developments to recent advancements. *Systems Science and Control Engineering*, 2(1), 602–609. https://doi.org/10.1080/21642583.2014.956265

Fix, E., & Hodges, J. L. (1951). Discriminatory Analysis. Nonparametric Discrimination: Consistency Properties, Report. USAF School of Aviation Medicine, Randolph Field, Texas.

Fletcher, R. (1987). Practical Methods of Optimization; (2nd Ed.). USA: Wiley-Interscience.

Friedman, J. H., Bentley, J. L., & Finkel, R. A. (1977). An Algorithm for Finding Best Matches in Logarithmic Expected Time. *ACM Transactions on Mathematical Software*, 3(3), 209–226. https://doi.org/10.1145/355744.355745

Gauch, M., Mai, J., & Lin, J. (2021). The proper care and feeding of CAMELS: How limited training data affects streamflow prediction. *Environmental Modelling and Software*, 135, 104926. https://doi.org/10.1016/j.envsoft.2020.104926

Genuer, R., Poggi, J. M., & Tuleau-Malot, C. (2010). Variable selection using random forests. *Pattern Recognition Letters*, 31(14), 2225–2236. https://doi.org/10.1016/j.patrec.2010.03.014

Ghosal, I., & Hooker, G. (2020). Boosting Random Forests to Reduce Bias; One-Step Boosted Forest and Its Variance Estimate. *Journal of Computational and Graphical Statistics*, 0(0), 1–30. https://doi.org/10.1080/10618600.2020.1820345

Griessinger, N., Seibert, J., Magnusson, J., & Jonas, T. (2016) Assessing the benefit of snow data assimilation for runoff modeling in Alpine catchments, *Hydrology and Earth System Sciences*, 20, 3895–3905, https://doi.org/10.5194/hess-20-3895-2016.

Grömping, U. (2009). Variable importance assessment in regression: Linear regression versus random forest. *American Statistician*, 63(4), 308–319. https://doi.org/10.1198/tast.2009.08199

Gupta, H., Kling, H., Yilmaz, K., & Martinez, G.: Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling, *Journal of Hydrology,* 377, 80–91, https://doi.org/http://dx.doi.org/10.1016/j.jhydrol.2009.08.003, 2009.

Hadi, S. J., Abba, S. I., Sammen, S. S. H., Salih, S. Q., Al-Ansari, N., & Mundher Yaseen, Z. (2019). Non-linear input variable selection approach integrated with non-tuned data intelligence model for streamflow pattern simulation. *IEEE Access*, 7, 141533–141548. https://doi.org/10.1109/ACCESS.2019.2943515

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition (2nd ed.). New York: Springer New York. Retrieved from https://books.google.ca/books?id=tVIjmNS3Ob8C

Hecht-Nielsen, R. (1989). Neurocomputing. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc.

Hornik, K., Stinchcombe, M., & White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5), 359–366. https://doi.org/https://doi.org/10.1016/0893-6080(89)90020-8

Jordan, A., Krüger, F., & Lerch, S. (2019). Evaluating Probabilistic Forecasts with {scoringRules}. *Journal of Statistical Software*, 90(12), 1–37. https://doi.org/10.18637/jss.v090.i12

Karlsson, M., & Yakowitz, S. (1987). Rainfall-runoff forecasting methods, old and new. *Stochastic Hydrology and Hydraulics,* 1, 303.

Konapala, G., Kao, S.-C., Painter, S. L., and Lu, D. (2020) Machine learning assisted hybrid models can improve streamflow simulation in diverse catchments across the conterminous US, *Environmental Research Letters*, 15, 10, 104022, https://doi.org/10.1088/1748-9326/aba927.

Kuczera, G., Renard, B., Thyer, M., Kavetski D. (2010) There are no hydrological monsters, just models and observations with large uncertainties! *Hydrological Sciences Journal*, 55 (6) (2010), pp. 980-991, 10.1080/02626667.2010.504677

Kratzert, F., Klotz, D., Herrnegger, M., Sampson, A. K., Hochreiter, S., & Nearing, G. S. (2019). Toward Improved Predictions in Ungauged Basins: Exploiting the Power of Machine Learning. *Water Resources Research*, 55(12), 11344–11354. https://doi.org/10.1029/2019WR026065

Kuhn, M., From, C., Wing, J., Weston, S., Williams, A., Keefer, C., et al. (2019). caret: Classification and Regression Training. Retrieved from https://cran.r-project.org/package=caret

Lakhanpal, A., Sehgal, V., Maheswaran, R., Khosa, R., & Sridhar, V. (2017). A non-linear and non-stationary perspective for downscaling mean monthly temperature: a wavelet coupled second order Volterra model. *Stochastic Environmental Research and Risk Assessment*, 31(9), 2159–2181. https://doi.org/10.1007/s00477-017-1444-6

Lall, U., & Sharma, A. (1996). A Nearest Neighbor Bootstrap For Resampling Hydrologic Time Series. *Water Resources Research*, 32(3), 679–693. https://doi.org/10.1029/95WR02966

Lee, T., Ouarda, T. B. M. J., & Yoon, S. (2017). KNN-based local linear regression for the analysis and simulation of low flow extremes under climatic influence. *Climate Dynamics*, 49(9–10), 3493–3511. https://doi.org/10.1007/s00382-017-3525-0

Li, J., Wang, Z., Lai, C., & Zhang, Z. (2019). Tree-ring-width based streamflow reconstruction based on the random forest algorithm for the source region of the Yangtze River, China. *Catena*, 183, 104216. https://doi.org/10.1016/j.catena.2019.104216

Li, Y., Li, C., Li, M., & Liu, Z. (2019). Influence of variable selection and forest type on forest aboveground biomass estimation using machine learning algorithms. *Forests*, 10(12). https://doi.org/10.3390/F10121073

Liu, T., Moore, A. W., Gray, A., and Yang, K. (2004). An investigation of practical approximate nearest neighbor algorithms. Adv. Neural Inf. Process. Syst., 17, 825–832.

Jiang, Z., Rashid, M. M., Johnson, F., & Sharma, A. (2020). A wavelet-based tool to modulate variance in predictors: an application to predicting drought anomalies. *Environmental Modelling and Software*, 104907. https://doi.org/https://doi.org/10.1016/j.envsoft.2020.104907

Maheswaran, R., & Khosa, R. (2012). Wavelet-Volterra coupled model for monthly stream flow forecasting. *Journal of Hydrology*, 450–451, 320–335. https://doi.org/10.1016/j.jhydrol.2012.04.017

Maheswaran, R., & Khosa, R. (2013). Long term forecasting of groundwater levels with evidence of non-stationary and nonlinear characteristics. *Computers and Geosciences*, 52, 422–436. https://doi.org/10.1016/j.cageo.2012.09.030

Mantovan, P., Todini, E. (2006) Hydrological forecasting uncertainty assessment: Incoherence of the GLUE methodology, *Journal of Hydrology*, 330, 1–2, 368-381, https://doi.org/10.1016/j.jhydrol.2006.04.046.

May, R. J., Maier, H. R., Dandy, G. C., & Fernando, T. M. K. G. (2008). Non-linear variable selection for artificial neural networks using partial mutual information. *Environmental Modelling and Software*, 23(10–11), 1312–1326. https://doi.org/10.1016/j.envsoft.2008.03.007

McInerney, D., Thyer, M., Kavetski, D., Lerat, J., & Kuczera, G. (2017), Improving probabilistic prediction of daily streamflow by identifying Pareto optimal approaches for modeling heteroscedastic residual errors, *Water Resources Research*, 53, 2199– 2239, doi:10.1002/2016WR019168.

McMillan, H., Krueger, T., Freer J. (2012) Benchmarking observational uncertainties for hydrology: rainfall, river discharge and water quality**,** *Hydrological Processes*, 26 (2012), pp. 4078-4111, 10.1002/hyp.9384

Mehta, P., Bukov, M., Wang, C. H., Day, A. G. R., Richardson, C., Fisher, C. K., & Schwab, D. J. (2019). A high-bias, low-variance introduction to Machine Learning for physicists. *Physics Reports*, 810, 1–124. https://doi.org/10.1016/j.physrep.2019.03.001

Mitchell, T. M. (1998). Machine learning, McGraw-Hill, New York.

Montanari, A., & Koutsoyiannis, D. (2012), A blueprint for process- based modeling of uncertain hydrological systems, *Water Resources Research,* 48, W09555, doi:10.1029/2011WR011412.

Müftüoğlu, R. F. (1984) New models for nonlinear catchment analysis, *Journal of Hydrology*, 73, 3–4, 335-357, https://doi.org/10.1016/0022-1694(84)90007-6.

Müftüoğlu, R. F. (1991) Monthly runoff generation by non-linear models, *Journal of Hydrology*, 125, 3–4, 277-291, https://doi.org/10.1016/0022-1694(91)90033-E.

Nearing, G. S., Kratzert, F., Sampson, A. K., Pelissier, C. S., Klotz, D., Frame, J. M., et al. (2020). What Role Does Hydrological Science Play in the Age of Machine Learning? *Water Resources Research*, e2020WR028091. https://doi.org/https://doi.org/10.1029/2020WR028091

Ni, L., Wang, D., Wu, J., Wang, Y., Tao, Y., Zhang, J., & Liu, J. (2020). Streamflow forecasting using extreme gradient boosting model coupled with Gaussian mixture model. *Journal of Hydrology*, 586, 124901. https://doi.org/10.1016/j.jhydrol.2020.124901

Padiyedath Gopalan, S., Kawamura, A., Amaguchi, H., Takasaki, T., Azhikodan, G. (2019) A bootstrap approach for the parameter uncertainty of an urban-specific rainfall-runoff model, *Journal of Hydrology*, 579, 124195. https://doi.org/10.1016/j.jhydrol.2019.124195.

Papacharalampous, G. A., & Tyralis, H. (2018). Evaluation of random forests and Prophet for daily streamflow forecasting. *Advances in Geosciences*, 45(2015), 201–208. https://doi.org/10.5194/adgeo-45-201-2018

Papacharalampous, G., Tyralis, H., Koutsoyiannis, D., & Montanari, A. (2020). Quantification of predictive uncertainty in hydrological modelling by harnessing the wisdom of the crowd: A large-sample experiment at monthly timescale. *Advances in Water Resources*, 136, 103470. https://doi.org/10.1016/j.advwatres.2019.103470

Papacharalampous, G., Tyralis, H., Langousis, A., Jayawardena, A. W., Sivakumar, B., Mamassis, N., et al. (2019a). Probabilistic Hydrological Post-Processing at Scale: Why and How to Apply Machine-Learning Quantile Regression Algorithms. *Water*, 11(10). https://doi.org/10.3390/w11102126

Papacharalampous, G., Tyralis, H., & Koutsoyiannis, D. (2019b). Comparison of stochastic and machine learning methods for multi-step ahead forecasting of hydrological processes. *Stochastic Environmental Research and Risk Assessment*, 6(d). https://doi.org/10.1007/s00477-018-1638-6

Pathy, A., Meher, S., & P, B. (2020). Predicting algal biochar yield using eXtreme Gradient Boosting (XGB) algorithm of machine learning methods. *Algal Research*, 50, 102006. https://doi.org/10.1016/j.algal.2020.102006

Pham, Q. B., Yang, T. C., Kuo, C. M., Tseng, H. W., & Yu, P. S. (2019). Combing random forest and least square support vector regression for improving extreme rainfall downscaling. *Water*, 11(3). https://doi.org/10.3390/w11030451

Pianosi, F., & Raso, L. (2012), Dynamic modeling of predictive uncertainty by regression on absolute errors, *Water Resources Research*, 48, W03516, doi:10.1029/2011WR010603.

Piryonesi, S. M., & El-Diraby, T. E. (2020). Role of Data Analytics in Infrastructure Asset Management: Overcoming Data Size and Quality Problems. *Journal of Transportation Engineering, Part B: Pavements*, 146(2), 04020022. https://doi.org/10.1061/jpeodx.0000175

Piotrowski, A.P., Napiorkowski, M. J., Napiorkowski, J. J., Osuch, M. & Kundzewicz, Z. W. (2017) Are modern metaheuristics successful in calibrating simple conceptual rainfall–runoff models? Hydrological Sciences Journal, 62:4, 606-625, DOI: 10.1080/02626667.2016.1234712.

Prasad, R., Deo, R. C., Li, Y., & Maraseni, T. (2018). Ensemble committee-based data intelligent approach for generating soil moisture forecasts with multivariate hydro-meteorological predictors. *Soil and Tillage Research*, 181, 63–81. https://doi.org/https://doi.org/10.1016/j.still.2018.03.021

Prasad, R., Deo, R. C., Li, Y., & Maraseni, T. (2019). Weekly soil moisture forecasting with multivariate sequential, ensemble empirical mode decomposition and Boruta-random forest hybridizer algorithm approach. *Catena*, 177, 149–166. https://doi.org/10.1016/j.catena.2019.02.012

Probst, P., & Boulesteix, A. L. (2018). To tune or not to tune the number of trees in random forest. *Journal of Machine Learning Research*, 18, 1–8.

Probst, P., Wright, M. N., & Boulesteix, A. L. (2019). Hyperparameters and tuning strategies for random forest. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 9(3), 1–15. https://doi.org/10.1002/widm.1301

Quilty, J., Adamowski, J., Khalil, B., & Rathinasamy, M. (2016). Bootstrap rank-ordered conditional mutual information (broCMI): A nonlinear input variable selection method for water resources modeling. *Water Resources Research*, 52(3), 2299–2326. https://doi.org/10.1002/2015WR016959

Quilty, J., & Adamowski, J. (2018). Addressing the incorrect usage of wavelet-based hydrological and water resources forecasting models for real-world applications with best practices and a new forecasting framework. *Journal of Hydrology*, 563, 336–353. https://doi.org/https://doi.org/10.1016/j.jhydrol.2018.05.003

Quilty, J., Adamowski, J., & Boucher, M.- A. (2019). A stochastic data- driven ensemble forecasting framework for water resources: A case study using ensemble members derived from a database of deterministic wavelet- based models. *Water Resources Research*, 55, 175– 202. https://doi.org/10.1029/2018WR023205.

Quilty, J., & Adamowski, J. (2020). A stochastic wavelet-based data-driven framework for forecasting uncertain multiscale hydrological and water resources processes. *Environmental Modelling & Software*, 130, 104718, https://doi.org/https://doi.org/10.1016/j.envsoft.2020.104718

Rahman, A. T. M. S., Hosono, T., Quilty, J. M., Das, J., & Basak, A. (2020). Multiscale groundwater level forecasting: Coupling new machine learning approaches with wavelet transforms. *Advances in Water Resources*, 141, 103595. https://doi.org/https://doi.org/10.1016/j.advwatres.2020.103595

Rajurkar, M.P., Kothyari, U.C., Chaube, U.C. (2004) Modeling of the daily rainfall-runoff relationship with artificial neural network, *Journal of Hydrology*, 285, 1–4, 96-113, https://doi.org/10.1016/j.jhydrol.2003.08.011.

Renard, B., Kavetski, D., Leblois, E., Thyer, M., Kuczera, G., & Franks, S. W. (2011), Toward a reliable decomposition of predictive uncertainty in hydrological modeling: Characterizing rainfall errors using conditional simulation, *Water Resources Research,*47, W11516, doi:10.1029/2011WR010643.

Ribeiro, G. T., Sauer, J. G., Fraccanabbia, N., Mariani, V. C., & dos Santos Coelho, L. (2020). Bayesian optimized echo state network applied to short-term load forecasting. *Energies*, 13(9). https://doi.org/10.3390/en13092390

Ripley, B. D. (1996). Pattern Recognition and Neural Networks. Cambridge University Press. https://doi.org/10.1017/CBO9780511812651.

Schaefli, B., Talamba, D., B., Musy, A. (2007) Quantifying hydrological modeling errors through a mixture of normal distributions, *Journal of Hydrology*, 332, 3–4, 303-315, https://doi.org/10.1016/j.jhydrol.2006.07.005.

Sehgal, V., Lakhanpal, A., Maheswaran, R., Khosa, R., & Sridhar, V. (2018). Application of multi-scale wavelet entropy and multi-resolution Volterra models for climatic downscaling. *Journal of Hydrology*, 556, 1078–1095. https://doi.org/https://doi.org/10.1016/j.jhydrol.2016.10.048

Seibert, J. (2000) Multi-criteria calibration of a conceptual runoff model using a genetic algorithm, *Hydrology and Earth System Sciences*, 4, 215–224, https://doi.org/10.5194/hess-4-215-2000.

Seibert, J., Vis, M. J. P. (2012) Teaching hydrological modeling with a user-friendly catchment-runoff-model software package, *Hydrology and Earth System Sciences*, 16, 3315–3325, https://doi.org/10.5194/hess-16-3315-2012.

Senent-Aparicio, J., Jimeno-Sáez, P., Bueno-Crespo, A., Pérez-Sánchez, P., Pulido-Velázquez, D. (2019) Coupling machine-learning techniques with SWAT model for instantaneous peak flow prediction, *Biosystems Engineering*, 177, 67-77, https://doi.org/10.1016/j.biosystemseng.2018.04.022.

Sharma, A., & Mehrotra, R. (2014). An information theoretic alternative to model a natural system using observational information alone. *Water Resources Research*, 50(1), 650–660. https://doi.org/10.1002/2013WR013845

Shen, C. (2018). A Transdisciplinary Review of Deep Learning Research and Its Relevance for Water Resources Scientists. *Water Resources Research*, 54. https://doi.org/10.1029/2018WR022643

Shortridge, J. E., Guikema, S. D., & Zaitchik, B. F. (2016). Machine learning methods for empirical streamflow simulation: A comparison of model accuracy, interpretability, and uncertainty in seasonal watersheds. *Hydrology and Earth System Sciences*, 20(7), 2611–2628. https://doi.org/10.5194/hess-20-2611-2016

Sikorska, A. E. & Renard, B.: Calibrating a hydrological model in stage space to account for rating curve uncertainties: general framework and key challenges, *Advances in Water Resources*, 105, 51–66, https://doi.org/110.1016/j.advwatres.2017.04.011, 2017.

Sikorska, A. & Seibert, J.: Appropriate temporal resolution of precipitation data for discharge modelling in pre-alpine catchments, *Hydrological Sciences Journal,* 61, 1–16, https://doi.org/10.1080/02626667.2017.1410279, 2018.

Sikorska, A.E., Montanari, A., & Koutsoyiannis, D., 2015. Estimating the uncertainty of hydrological predictions through data-driven resampling techniques. *Journal of Hydrologic Engineering*, 20, A4014009.https://doi.org/10.1061/(ASCE)HE.1943-5584.

Sikorska, A.E., Scheidegger, A., Banasik, K., & Rieckermann, J. (2012) Bayesian uncertainty assessment of flood predictions in ungauged urban basins for conceptual rainfall-runoff models, *Hydrology and Earth System Sciences*, 16, 1221-1236, doi:10.5194/hess-16-1221-2012.

Sikorska-Senoner A. E. & Seibert J. (2020) Flood-type trend analysis for alpine catchments, *Hydrological Sciences Journal*, 65:8, 1281-1299, DOI: 10.1080/02626667.2020.1749761.

Sikorska-Senoner, A. E., Schaefli, B., & Seibert, J. (2020) Downsizing parameter ensembles for simulations of rare floods, *Natural Hazards and Earth System Sciences*, 20, 3521–3549, https://doi.org/10.5194/nhess-20-3521-2020, 2020.

Snoek, J., Larochelle, H., & Adams, R. P. (2012). Practical bayesian optimization of machine learning algorithms. *Advances in Neural Information Processing Systems*, 25, 2951–2959.

Solomatine, D. P., Ostfeld, A. (2008) Data-driven modelling: some past experiences and new approaches. Journal of Hydroinformatics 1 January 2008; 10 (1): 3–22. doi:https://doi.org/10.2166/hydro.2008.015.

Solomatine, D.P., & Xue, Y. (2004) M5 model trees and neural networks: application to flood forecasting in the upper reach of the Huai River in China. *Journal of Hydrologic Engineering*, 9 (6),491–501.

Sorooshian, S., & Dracup, J. A. (1980), Stochastic parameter estimation procedures for hydrologie rainfall- runoff models: Correlated and heteroscedastic error cases, *Water Resources Research*, 16(2), 430– 442, doi:10.1029/WR016i002p00430.

Smith, T., Marshall, L., & Sharma, A. (2015) Modeling residual hydrologic errors with Bayesian inference, *Journal of Hydrology*, 528, 29-37, https://doi.org/10.1016/j.jhydrol.2015.05.051.

Sun, W., & Trevor, B. (2017). Combining k-nearest-neighbor models for annual peak breakup flow forecasting. *Cold Regions Science and Technology*, 143, 59–69. https://doi.org/10.1016/j.coldregions.2017.08.009

Suryanarayana, C., Sudheer, C., Mahammood, V., & Panigrahi, B.K. (2014). An integrated wavelet-support vector machine for groundwater level prediction in Visakhapatnam, India. *Neurocomputing*, 145, 324–335.

Tencaliec, P., Favre, A.-C., Prieur, C., Mathevet, T. (2015) Reconstruction of missing daily streamflow data using dynamic regression models. *Water Resources Research*, 51, 9447-9463. https://doi.org/10.1002/2015WR017399.

Teweldebrhan, A. T., Schuler, T. V, Burkhart, J. F., & Hjorth-Jensen, M. (2020). Coupled machine learning and the limits of acceptability approach applied in parameter identification for a distributed hydrological model. *Hydrology and Earth System Sciences*, 24(9), 4641–4658. https://doi.org/10.5194/hess-24-4641-2020

Tongal, H., & Booij M. J. (2018) Simulation and forecasting of streamflows using machine learning models coupled with base flow separation, *Journal of Hydrology*, 564, 266-282, https://doi.org/10.1016/j.jhydrol.2018.07.004.

Tsimpiris, A., Vlachos, I., & Kugiumtzis, D. (2012). Nearest neighbor estimate of conditional mutual information in feature selection. *Expert Systems with Applications*, 39(16), 12697–12708. https://doi.org/10.1016/j.eswa.2012.05.014

Tyralis, H., & Papacharalampous, G. (2017). Variable selection in time series forecasting using random forests. *Algorithms*, 10(4). https://doi.org/10.3390/a10040114

Tyralis, H., Papacharalampous, G., Burnetas, A., & Langousis, A. (2019a). Hydrological post-processing using stacked generalization of quantile regression algorithms: Large-scale application over CONUS. *Journal of Hydrology*, 577, 123957. https://doi.org/10.1016/j.jhydrol.2019.123957

Tyralis, H., Papacharalampous, G., & Langousis, A. (2019b). A Brief Review of Random Forests for Water Scientists and Practitioners and Their Recent History in Water Resources. *Water*, 11(5). https://doi.org/10.3390/w11050910

Tyralis, H., Papacharalampous, G., & Langousis, A. (2020). Super ensemble learning for daily streamflow forecasting: large-scale demonstration and comparison with multiple machine learning algorithms. *Neural Computing and Applications*, 3. https://doi.org/10.1007/s00521-020-05172-3

Vanschoren, J., Blockeel, H., Pfahringer, B., & Holmes, G. (2012). Experiment databases: A new way to share, organize and learn from experiments. *Machine Learning*, 87(2), 127–158. https://doi.org/10.1007/s10994-011-5277-0

Venables, W. N., & Ripley, B. D. (2002). Modern Applied Statistics with S (Fourth). New York: Springer. Retrieved from http://www.stats.ox.ac.uk/pub/MASS4

Villalba, G., Liang, X., Liang, Y. (2021) Selection of multiple donor gauges via Graphical Lasso for estimation of daily streamflow time series. *Water Resources Research*, e2020WR028936. https://doi.org/10.1029/2020WR028936.

Vlachos, I., & Kugiumtzis, D. (2010). Nonuniform state-space reconstruction and coupling detection. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, 82(1), 1–16. https://doi.org/10.1103/PhysRevE.82.016207

Wang, J., Bao, W., Gao, Q., Si, W., Sun, Y. (2021) Coupling the Xinanjiang model and wavelet-based random forests method for improved daily streamflow simulation. *Journal of Hydroinformatics*, jh2021111. doi:https://doi.org/10.2166/hydro.2021.111.

Wani, O., Beckers, J. V. L., Weerts, A. H., & Solomatine, D. P. (2017). Residual uncertainty estimation using instance-based learning with applications to hydrologic forecasting. *Hydrology and Earth System Sciences*, 21(8), 4021–4036. https://doi.org/10.5194/hess-21-4021-2017

Westerberg, I. K., Sikorska-Senoner, A. E., Viviroli, D., Vis, M., & Seibert, J. (2020) Hydrological model calibration with uncertain discharge data, *Hydrological Sciences Journal*, DOI: 10.1080/02626667.2020.1735638

Wilson, S. (2019). ParBayesianOptimization: Parallel Bayesian Optimization of Hyperparameters. Retrieved from https://cran.r-project.org/package=ParBayesianOptimization

Wright, M. N., & Ziegler, A. (2017). ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. *Journal of Statistical Software*, 1(1). Retrieved from https://www.jstatsoft.org/v077/i01

Wu, T., & Kareem, A. (2014). Simulation of nonlinear bridge aerodynamics: A sparse third-order Volterra model. *Journal of Sound and Vibration*, 333(1), 178–188. https://doi.org/10.1016/j.jsv.2013.09.003

Wu, W., Dandy, G. C., & Maier, H. R. (2014). Protocol for developing ANN models and its application to the assessment of the quality of the ANN model development process in drinking water quality modelling. *Environmental Modelling & Software*, 54, 108–127. https://doi.org/10.1016/j.envsoft.2013.12.016

Wu, R., Yang, L., Chen, C., Ahmad, S., Dascalu, S. M., & Harris Jr., F. C. (2019) MELPF version 1: Modeling Error Learning based Post-Processor Framework for Hydrologic Models Accuracy Improvement, *Geoscientific Model Development*, 12, 4115-4131, https://doi.org/10.5194/gmd-12-4115-2019.

Xiong, L., Wan, M., Wei, X., & O'Connor, K. M. (2009). Indices for assessing the prediction bounds of hydrological models and application by generalised likelihood uncertainty estimation. *Hydrological Sciences Journal*, 54(5), 852–871. https://doi.org/10.1623/hysj.54.5.852

Yang, J., Reichert, P., Abbaspour, K. C., & Yang, H. (2007) Hydrological modelling of the Chaohe Basin in China: Statistical model formulation and Bayesian inference, *Journal of Hydrology*, 340, 3–4, 167-182, https://doi.org/10.1016/j.jhydrol.2007.04.006.

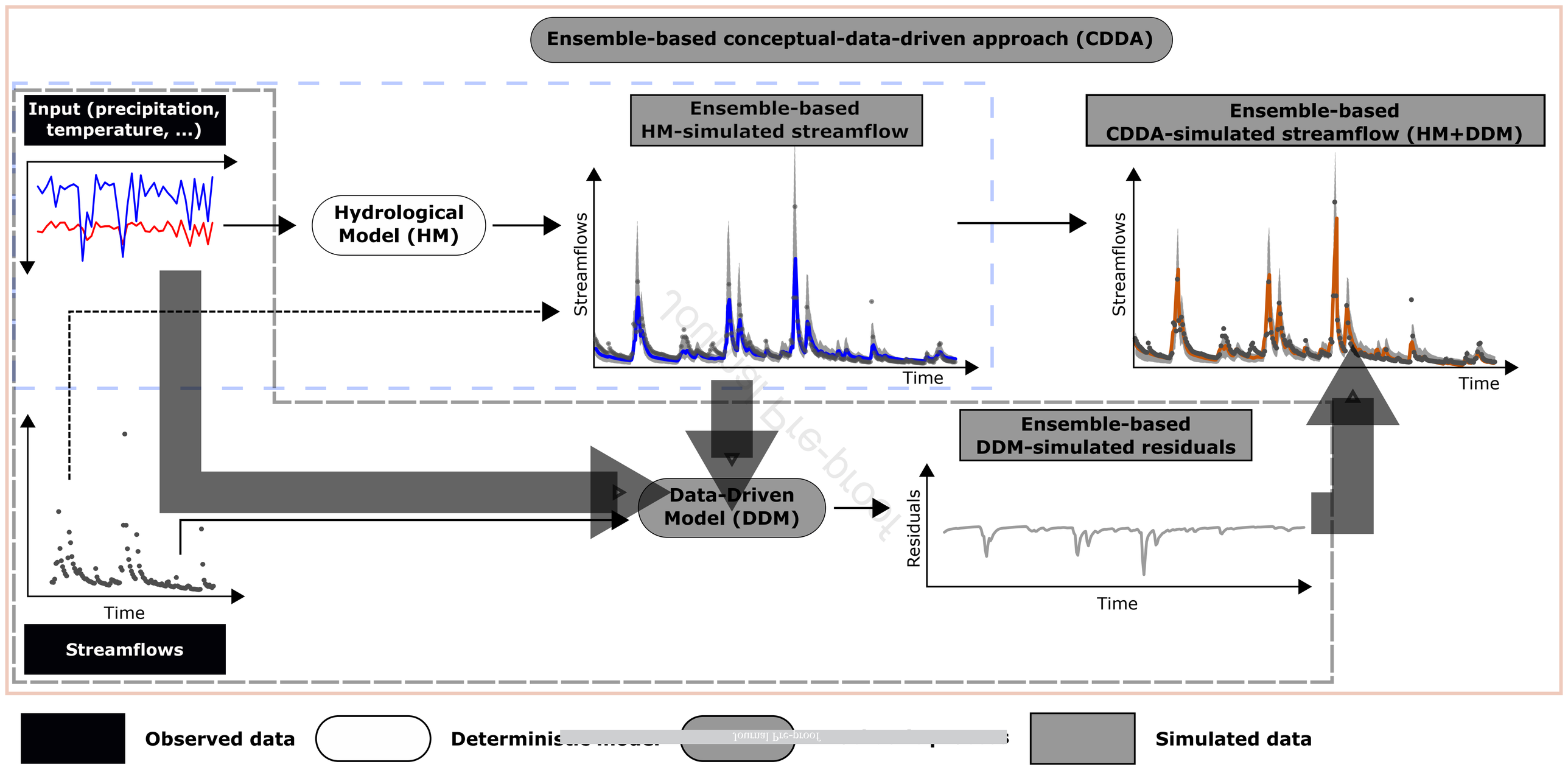
Yang, S., Yang, D., Chen, J., Santisirisomboon, J., Lu, W., & Zhao, B. (2020a) A physical process and machine learning combined hydrological model for daily streamflow simulations of large watersheds with limited observation data, *Journal of Hydrology*, 590, 125206, https://doi.org/10.1016/j.jhydrol.2020.125206.

Yang, Z., Zhang, A., & Sudjianto, A. (2020b). Enhancing Explainability of Neural Networks Through Architecture Constraints. *IEEE Transactions on Neural Networks and Learning Systems*, 1–12. https://doi.org/10.1109/TNNLS.2020.3007259

Zambrano-Bigiarini, M. (2017). hydroGOF: Goodness-of-fit functions for comparison of simulated and observed hydrological time series. https://doi.org/10.5281/zenodo.840087

Zhang, Y., Zhang, R., Ma, Q., Wang, Y., Wang, Q., Huang, Z., & Huang, L. (2020). A feature selection and multi-model fusion-based approach of predicting air quality. *ISA Transactions*, 100, 210–220. https://doi.org/10.1016/j.isatra.2019.11.023

Zuo, G., Luo, J., Wang, N., Lian, Y., & He, X. (2020). Two-stage variational mode decomposition and support vector regression for streamflow forecasting. *Hydrology and Earth System Sciences*, 24(11), 5491–5518. https://doi.org/10.5194/hess-24-5491-2020


Ensemble-based conceptual-data-driven approach (CDDA)
Input (precipitation, temperature, ...)
Hydrological Model (HM)
Ensemble-based HM-simulated streamflow
Streamflows
Time
Ensemble-based CDDA-simulated streamflow (HM+DDM)
Streamflows
Time
Ensemble-based DDM-simulated residuals
Residuals
Time
Data-Driven Model (DDM)
Time
Streamflows
Observed data
Simulated data

A novel ensemble-based conceptual-data-driven approach for improved streamflow simulations

Anna E. Sikorska-Senoner[1] and John M. Quilty[2]

Highlights:

- Conceptual-data-driven approach (CDDA) proposed for ensemble streamflow simulation
- The CDDA couples a data-driven model (DDM) and a hydrological model (HM)
- Eight DDMs are explored as a potential predictor of the HM residual ensemble
- CDDA improves the mean continuous ranked probability score vs. standalone HM
- eXtreme Gradient Boosting and Random Forests are recommended to modelHM residuals

Declaration of interests

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: